Computerlinguistik und Künstliche Intelligenz

Datengesteuerte maschinelle Übersetzung mit flachen Analysestrukturen

MAGISTERARBEIT
ZUR ERLANGUNG DES
MAGISTER ARTIUM

im Fachbereich
Sprach- und Literaturwissenschaft
der Universität Osnabrück

vorgelegt von:

Joachim Wagner
aus Nordenham (Geburtsort)

2003

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

1.1 Computer Aided Translation

Um das Thema meiner Magisterarbeit auch für Leser ohne Kenntnis der Begriffe verständlich zu machen, die ich im Titel verwendet habe, möchte ich zuerst erläutern, in welchem Kontext ein Translation Memory verwendet wird. Ich werde dabei etwas weiter ausholen, um das Thema besser von verwandten Themen, wie z.B. Example Based Machine Translation, abgrenzen zu können. Für das Verständnis der verschiedenen Ansätze ist es zudem hilfreich, die verschiedenen Anwendungsgebiete vor Augen zu haben, aus denen sich unterschiedlichen Anforderungen ableiten lassen.

1.1.1 Anwendungsgebiete

Das Anfertigen von Übersetzungen hat durch die Globalisierung und Internationalisierung von Wirtschaft, Politik und Kultur in den letzten Jahrzehnten stark an Bedeutung gewonnen. Waren, die in vielen Ländern verkauft werden, müssen an die Vorschriften der Zielländer und an die Kundenwünsche angepasst werden. Hierzu zählt insbesondere, dass die aufgedruckten oder beigelegten Texte, die z.B. wichtige Hinweise zur Handhabung enthalten, in die Sprachen der Zielländer übersetzt werden. Sprachbarieren müssen im Wirtschaftsleben nicht nur bei Kunden- und Geschäftsbeziehungen überwunden werden. Auch innerhalb internationaler Unternehmen besteht Bedarf an Übersetzungen. Insbesondere nach einer Fusion von Partnern aus unterschiedlichen Sprachregionen stellt sich das Problem, wie die verschiedensprachigen Mitarbeiter Zugang zum in Dokumente abgelegten Wissen des neuen Unternehmens erhalten. Moderne Sprachtechnologie ermöglicht es zwar, zu einer Fragestellung relevante Dokumente über Sprachgrenzen hinweg ausfindig zu machen. Anschließend wird aber eine Übersetzung benötigt, sofern die Sprachkenntnisse der Mitarbeiter nicht ausreichen. In der Regel wird dies eine vom Computer erstellte Übersetzung sein, die es erlaubt, den Inhalt grob zu verstehen. Leider sind durch Computersoftware erstellte Übersetzungen häufig[1] un- oder missverständlich, da es derzeit noch sehr schwierig ist, Satzbau, Wortbeziehungen und Fachbegriffe inhaltlich angemessen zu interpretieren. Daher muss, wenn es auf hohe Genauigkeit ankommt, ein professioneller Übersetzer beauftragt werden. Dem Übersetzer wird dabei nicht nur Sprach-, sondern auch Fachkompetenz abverlangt.

[1] Die Qualität hängt von vielen Faktoren ab und zwar nicht nur von den beteiligten Sprachen und dem machinellen Verfahren. Das Genre, das verwendete Vokabular, der Satzbau und andere stilistische Eigenschaften des Ausgangstextes beeinflussen den Erfolg der maschinellen Übersetzung erheblich.

Ähnlich verhält es sich im Bereich der Politik. Z. B. wird von der Öffentlichkeit kaum beachtet, dass für die Europäischen Kommissionen viele Fachaufsätze, Pressematerialien und Dokumente zu Gesetzen und Reden zwischen den elf offiziellen EU-Sprachen übersetzt werden. Wie in der Wirtschaft werden je nach Verwendung des Dokuments unterschiedliche Anforderungen an die Übersetzung gestellt. Für die Außendarstellung ist es wichtig, dass Begriffe einheitlich wiedergegeben werden, und zwar nicht nur innerhalb eines Textes, sondern auch im Bezug auf zurückliegende Veröffentlichungen. Bei Verordnungen und Richtlinien tritt u.a. hinzu, dass die Textstruktur soweit erhalten bleiben muss, dass Verweise auf einzelne Absätze und Sätze auch in der Übersetzung korrekt sind.

Schließlich sei noch die Kultur betrachtet. Auch wenn die Entscheidung, ein Buch in eine andere Sprache zu übersetzen, meistens wohl von wirtschaftlichen Gewinnerwartungen bestimmt wird, kann der Einfluss der vielen angebotenen Bücher anderssprachiger Autoren auf die eigene Kultur nicht geleugnet werden.[2] Die verschiedenen Kategorien von Büchern, wie z.B. Biografien, Krimis und Reiseführer, stellen unterschiedliche Anforderungen an den Überetzer. Das gleiche gilt für Filme. Sie werden nicht etwa von Dolmetschern simultan übersetzt. Synchronsprecher und Übersetzer sind in der Regel verschiedene Personen. Trotzdem unterscheidet sich das Übersetzen eines Filmskripts von dem eines Romans. Der Text muss so gestaltet werden, dass der Sprecher ihn lippen- und atmungssynchron sprechen kann. Dies beschränkt u.a. die Wortwahl und die Länge der Übersetzung.

Die obigen Beispiele aus verschiedenen Bereichen zeigen, dass Übersetzen mehr verlangt als Vokabular und Grammatik der beteiligten Sprachen zu beherrschen. Der Übersetzer muss die Funktion des Textes, die kulturellen Unterschiede zwischen der alten und der neuen Zielgruppe und die Eigenheiten der jeweiligen Fachsprache und Textgattung beachten. Zu seinen Aufgaben gehört es daher, sich in den Hintergrund einzuarbeiten, Vorschläge für inhaltliche Änderungen zu erarbeiten und sich mit dem Fachvokabular vertraut zu machen.[3] Bei den notwendigen Recherchen helfen Nachschlagewerke und Dienstleister. Zugriffsmöglichkeiten auf solche Informationsquellen gehören zum modernen Computerarbeitsplatz eines Übersetzers genauso wie Software, die bei der eigentlichen Übersetzungsarbeit hilft.

1.1.2 Software-Werkzeuge

Eine ganze Reihe von Software-Werkzeugen stehen dem Übersetzer heute zur Verfügung. Zum einen sind dies elektronische (Fach-) Wörterbücher, die das Nachschlagen beschleunigen, Platz auf dem Schreibtisch sparen, und die Einträge übersichtlicher präsentieren können. Im Gegensatz zu einem normalen Wörterbuchbenutzer ist es für einen Übersetzer besonders wichtig, eigene Einträge z.B. zu der speziellen Terminologie, die in den Texten eines Auftraggebers vorkommt, erstellen zu können. Soll umfangreiches Material übersetzt werden, dann erleichtert es eine solche kundenspezifische Terminologie-Zusammenstellung, die Begriffe korrekt und einheitlich zu übersetzen.[4] Häufig werden auch einsprachige Beschreibungen der Terminologie genutzt, die der Auftraggeber zur Verfügung stellt, oder die gewünschten Übersetzungen der Begriffe können aus bereits übersetzen

[2] Es soll hier aber auch nicht um den kulturellen Wert dieser Bücher gehen, sondern darum, die Allgegenwertigkeit von Übersetzungen in unserer Zeit und die Breite der Anforderungen an Übersetzungen zu verdeutlichen.

[3] Viele Übersetzungsdienstleister bieten neben Übersetzungsleistungen auch das Anfertigen von Zusammenfassungen und das Überarbeiten von Manuskripten an.

[4] Dass das Wörterbuch und die Terminologiepflege i.d.R. Produkte verschiedener Hersteller sind, muss hier nicht weiter interessieren. In der Praxis bedeutet das lediglich, dass der Benutzer vor dem Nachschlagen entscheiden muss, welches Verzeichnis er wählt.

Texten extrahiert werden. Es wird bereits Software angeboten, die diese Extraktion automatisch durchführt. Allerdings sind die Terminologieextraktion und die Identifikation der entsprechenden Übersetzung in gegebenen Paaren von Ausgangstexten und ihren Übersetzungen aktive Forschungsgebiete.

Ein Terminologie-Manager kann sich in der Art der Benutzung von einem Wörterbuch unterscheiden. Da Terminologie innerhalb eines Projekt und häufig darüber hinaus einheitlich übersetzt wird, kann er dem Übersetzer unaufgefordert auf die Übersetzung hinweisen.

Ein weiteres Werkzeug ist der Concordancer. Er zeigt in verschiedenen Darstellungen alle mit einer Eingabe übereinstimmenden Textstellen an. Bereits wenn einsprachige Texte in beiden an der Übersetzung beteiligten Sprachen vorliegen, können Unterschiede in der Verwendung eines Begriffs und seiner (vermuteten) Übersetzung untersucht werden. Um zu wertvollen Erkenntnissen zu gelangen, reicht es völlig aus, dass die Texte aus vergleichbaren Bereichen kommen. Man spricht hier auch von Vergleichskorpora oder vergleichbaren Korpora (comparable corpora). Da normalerweise beim Übersetzen ein natürlich wirkender Text enstehen soll, wählt man auch für die Zielsprache Texte, die in dieser Sprache ursprünglich verfasst wurden. Bowker (1998) zeigt, dass einsprachiges Material dem Übersetzer helfen kann, den Ausgangstext besser zu verstehen und sich in der Zielsprache treffender auszudrücken. Bowker hat in seinem Experiment Testpersonen neben einen Concordancer auch zwei statistische Werkzeuge zur Verfügung gestellt. Das eine Werkzeug extrahiert auffällige Wortkombinationen (sogenannte Kollokationen, siehe Abschnitt 3.1.1). Es kann z.B. eine Rangliste der Wörter erstellen, die zusammen mit einem vorgegebenen Wort auftreten. Das andere statistische Werkzeug zeigt die Verteilung der Verwendungen von Ausdrücken im Textmaterial an und gibt damit einen Hinweis darauf, ob es sich um verbreitete Ausdrucksweise oder um spezielle einzelner Autoren handelt.

Wenn jedoch Unsicherheiten bestehen, ob alle in Frage kommenden Übersetzungen bekannt sind, oder wenn untersucht werden soll, unter welchen Bedingungen welche Übersetzung gewählt wird, dann werden Texte zusammen mit ihrer Übersetzung benötigt. Solches Material wird paralleles Korpus, bilinguales Korpus oder Bitext genannt. Je nach Anwendung ist es wichtig, dass nicht zu frei übersetzt wurde und dass die Übersetzungsrichtung einheitlich ist, d.h. dass Ausgangs- und Zielsprache nicht wechseln. Auch sind Texte problematisch, die aus einer dritten, nicht am Korpus beteiligten Sprache übersetzt wurden. Für die Arbeit eines Übersetzers sind die Ergebnisse früherer Übersetzungsbemühungen des gleichen Auftraggebers besonders aufschlussreich. Ein bilingualer Concordancer zeigt Textstellen zusammen mit ihrer Übersetzung an. Wahlweise können für eine oder beide Sprachseiten Wörter vorgegeben werden, die in den anzuzeigenden Stellen auftreten müssen. Hier übernimmt der Übersetzer Aufgaben, die eigentlich zu dem Arbeitsbereich eines Lexikographen gehören. Concordancer sind besonders hilfreich, wenn die Zielsprache der Übersetzung nicht die Muttersprache des Übersetzers ist. Es können Belege für Formulierungen gesucht und typische sprachliche Muster erkannt werden.

Wie bereits weiter oben erwähnt steht auch Software zur Verfügung, die eine Übersetzung automatisch erstellt. Zur maschinellen Übersetzung (machine translation, MT) sind einige populäre Irrtümer verbreitet, die solche Systeme in ein schlechtes Licht rücken. So sei MT grundsätzlich unbrauchbar, da sie den Sinn entstelle und zu viele Korrekturen erfordere. Richtig ist zwar, dass durch ein heutiges MT-System erstellte Übersetzungen grobe und sehr eigensinnige Mängel aufweisen. Welcher Anteil der Übersetzung unverständlich wird, hängt aber von den beteiligten Sprachen, dem benutzten MT-System und von den Eigenschaften des Ausgangstextes ab. Wenn bereits bei der Erstellung des Ausgangstextes auf einen einfachen Satzbau geachtet wurde, kann mit MT eine Rohübersetzung erstellt werden, deren Nachbearbeitungsaufwand geringer ist als der Aufwand einer manuellen

Übersetzung.[5] Das hängt natürlich auch von der Arbeitsweise des Übersetzers ab. MT-Systeme haben aber schon dadurch Berechtigung, dass Übersetzungsdienstleister mit ihnen eine schnelle Rohübersetzung anbieten können. Nicht jeder Auftraggeber benötigt eine sprachlich einwandfreie Übersetzung. Für viele Zwecke reicht eine Übersetzung aus, die es erlaubt, den Inhalt des Ausgangstextes zu erschließen.

MT heißt nicht zwangsläufig, dass ein Ausgangstext in das System eingegeben wird und ohne jede Benutzerinteraktion eine Übersetzung entsteht. Nach der Art der Interaktion werden zwei Strategien unterschieden: HAMT (human aided machine translation) und MAHT (machine aided human translation). Bei der vom Menschen unterstützten maschinellen Übersetzung (HAMT) stellt der Computer dem Benutzer Fragen, z.B. wenn es Unsicherheiten bei der Interpretation des Ausgangstextes gibt. Gerne gewähltes Beispiel ist hier die Anaphernresolution, d. h. das Finden des Bezugs eines Pronomens. Das System zeigt den Ausgangstext an, hebt das Pronomen und in Frage kommende Antezedenzien hervor und bitten den Benutzer, eine Entscheidung zu treffen. Dieser Art der Übersetzung hat den Nachteil, dass der Benutzer in eine passive Rolle gezwungen wird. Die Fragen sind zahlreich und häufig anspruchslos. Nicht jede Mehrdeutigkeit wird erkannt, sodass immer noch eine Nachbearbeitung der Übersetzung notwendig ist.

Im Gegensatz dazu übernimmt bei der maschinengestützten Übersetzung (MAHT) der Übersetzer die aktive Rolle.

Die Art, wie der Computer dem Übersetzer hilft, kann sehr unterschiedlich sein. Es gibt Systeme, die aufgrund des Ausgangstextes[6] während der Eingabe der Übersetzung Vorhersagen treffen, welches Wort gerade geschrieben werden soll. Da ein erfahrender Schreiber beim Maschinenschreiben auf den Bildschirm schaut, kann er die Vorhersage mit einem Tastendruck übernehmen und so die Schreibgeschwindigkeit erhöhen. Interessanter ist aber die Möglichkeit, bereits bevor das erste Zeichen eines Wortes eingegeben wurde Vorschläge für das nächste Wort zu erhalten. Mehrdeutigkeiten, die ein MT-System zu einer möglicherweise falschen Entscheidung zwingen, können hier offen bleiben.

Ein weiteres Hilfsmittel ist das Translation Memory, kurz TM, um das es in der vorliegenden Arbeit geht. Der Begriff TM steht sowohl für das Hilfsmittel als auch für das zweisprachige Textmaterial, auf das es zugreift. Mit einem bilingualen Concordancer hat ein TM aber nicht viel gemeinsam. Der Zweck eines TMs ist, Sätze (oder andere Texteinheiten), die schon einmal übersetzt wurden, nicht erneut übersetzen zu müssen, sondern die Übersetzung aus dem vorhandenen Textmaterial abrufen zu können, sodass während des Übersetzungsprozesses die bereits geleistete Übersetzungsarbeit genutzt werden kann. Ein TM sucht dazu eine passende Stelle im ausgangssprachlichen Material und identifiziert dann die Übersetzung in der zielsprachlichen Seite des Textmaterials. Das im TM gespeicherte Material wird daher auch als Referenzmaterial bezeichnet. Die Identifikation der Übersetzung erfordert, dass eine Sprachseite des Textmaterials Übersetzung der anderen Seite sein muss. Vergleichbarkeit der Texte reicht nicht aus.[7] Anders als bei einem bilingualen Concordancer werden nicht einzelne Wörter, sondern längere Einheiten, meistens ganze Sätze, abgefragt. Des Weiteren ist ein TM nicht als Recherchewerkzeug ausgelegt. Es tritt gewöhnlich von selbst in Aktion, bevor ein Satz übersetzt werden soll.

[5] Z.B. lassen verschiedene kanadische Einrichtungen Wetterberichte u.ä. durch das MT-System METEO von der Firma Chandioux (`http://www.chandioux.com/`) ins Französische übersetzen.

[6] Denkbar wäre auch, ein Vorhersagesystem zu entwickeln, dass monolingual arbeitet, d.h. nur die bisher geschriebene Übersetzung und Wissen über die Zielsprache nutzt.

[7] Man könnte sich auch ein System vorstellen, das mit lediglich vergleichbaren Texten arbeitet und anhand von Merkmalen des Ausgangssatzes einen Satz des Textmaterials als Übersetzungsvorschlag auswählt. Wenn die Menge der im Textmaterial vorzufindenen Kontexte des Satzes nicht zu den verwendeten Merkmalen gehört, dann benötigt man also nur Texte der Zielsprache. Im Prinzip läuft es dann auf ein MT-System hinaus, das nur sprachliche Ausgaben produziert, die wortwörtlich im Textmaterial belegt sind.

Beim Übersetzen von Bedienungsanleitungen, Handbüchern und anderen Texten, zu denen eine ältere Fassung bereits übersetzt wurde, kann ein Translation Memory (TM) helfen, Zeit zu sparen. Zu Sätzen, die wortwörtlich im Referenzmaterial vorhanden sind, kann die alte Übersetzung i.d.R. ohne Rückfragen übernommen werden. Der Übersetzer muss nur eingreifen, wenn zum Ausgangstext kein Referenzmaterial gefunden werden kann.

Zentrales Werkzeug für den Übersetzer ist jedoch ein spezielles Textverarbeitungsprogramm, das den bereits vorhandenen Ausgangstext besonders berücksichtigt. Das Anfertigen einer Übersetzung unterscheidet sich vom Verfassen eines neuen Textes insbesondere in folgenden Punkten. So können die Struktur und Formatierung des Ausgangstextes übernommen werden. Auch macht es Sinn, während des Schreibens die zugehörige Stelle im Ausgangstext fortlaufend auf dem Bildschirm anzuzeigen oder eine Möglichkeit anzubieten, auf Verlangen zu ihr zu springen. Dieses und viele andere Kleinigkeiten können die Produktivität erheblich steigern.

Ein weiterer wichtiger Aspekt ist das Zusammenspiel der einzelnen Komponenten. Z.B. darf es nicht zu umständlich sein, während des Schreibens mögliche Übersetzungen zu einem Wort des Ausgangstextes oder Synonyme eines gerade geschriebenen Wortes abzufragen. Ziel ist es, den Übersetzer bei seiner Arbeit so gut wie möglich durch den Computer zu unterstützen. Man spricht daher von computer-assisted translation (CAT). Der Begriff überschneidet sich mit machine-aided human translation (MAHT, siehe oben).

1.2 Zielsetzung

In dieser Arbeit möchte ich eine Idee aufgreifen, die mir mein Zweitbetreuer Helmar Gust im Anschluss an einen Vortrag vorstellte. Gewöhnliche Translation Memorys nutzen nur einen Satz aus dem Referenzmaterial. Zwar können sie dem Übersetzer alle Fundstellen im Referenzmaterial anzeigen. Aber letztendlich muss er einen Satz auswählen, dessen Übersetzung als Vorlage dienen soll. Das Zusammensetzen der Übersetzung aus verschiedenen Fundstellen wird von TM-Software aus guten Gründen nicht unterstützt.

Wenn man auf mehrere im Translation Memory gespeicherte Sätze, die nur teilweise mit dem zu übersetzenden Satz übereinstimmen, zurückgreifen möchte, um eine Rohübersetzung zu generieren, dann treten viele Probleme auf. Die Übersetzungen der übereinstimmenden Passagen der Referenzsätze müssen identifiziert und zu einem neuen Satz zusammengesetzt werden. Hierbei kann je nach Zielsprache die Reihenfolge der Teile eine Rolle spielen, und die Teile können aus verschiedenen Gründen nicht zusammen passen. Zum Beispiel kann die Übersetzung 'sprangen ... aus dem Zug' von '... hopped off the train' im Deutschen nur in der ersten und dritten Person Plural benutzt werden. Maschinell zu überprüfen, ob wie im Beispiel Person und Numerus abweichen, ist schwierig. Es ist aber auch nicht notwendig, da es für einen Übersetzer einfach ist, die Flexion anzupassen.

Die Idee ist nun, diese Probleme zu reduzieren, indem nur solche Referenzsätze herangezogen werden, deren syntaktische Struktur mit der des zu übersetzenden Satzes übereinstimmt. Die Struktur kann u. a. an der Abfolge der Wortarten erkannt werden. Ein sehr einfacher Ansatz könnte verlangen, dass die Wortarten vollständig übereinstimmen. Dann werden in der Regel[8] die syntaktischen Strukturen — angefangen von der Abfolge der einzelnen Satzteile bis hin zu der inneren Struktur der Phrasen — den gleichen Aufbau haben. Eventuell müssen für bestimmte Wortarten, z.B. Präpositionen und Verben, auch die Wörter bzw. Grundformen übereinstimmen, um unbrauchbare Referenzsätze auszu-

[8] Trotz gleicher Folge von Wortarten kann die syntaktische Struktur abweichen. Vergleiche z.B. 'Er sah den Mann mit dem Hut.' und 'Er sah das Reh mit dem Fernglas.'

schließen. Dies in Ansätzen zu untersuchen wird der zentrale Gegenstand der vorliegenden Magisterarbeit sein.

Ein weiteres Problem ist die Identifikation der Übersetzung von den Teilen der Referenzsätze, auf die zurückgegriffen werden soll. Hier bieten sich zwei grundsätzliche Vorgehensweisen an. Zum einen könnte man auf einen der verschiedenen bereits veröffentlichten Ansätze zurückgreifen. Problematisch ist, dass die meisten Ansätze Terminologie oder Phrasen aus größeren Korpora und nicht aus einzelnen Satzpaaren extrahieren. Alternativ könnte man den Ansatz aus dem Studienprojekt KoKS (Erpenbeck et al., 2002) verwenden, mit dem ich vertraut bin, da ich Mitglied dieses Projekts war. Die Ergebnisse des Studienprojekts zeigen aber, dass der Ansatz noch nicht ausgereift ist. Es treten viele falsche Zuordnungen auf.

In dieser Arbeit soll ein Ansatz mit Hilfe von Beispielen aus einem Deutsch-Englischen Übersetzungskorpus skizziert werden, der sich nur auf einfache linguistischer Werkzeuge, nämlich POS-Tagging und Lemmatisierung, und parallele Korpora stützt. Dies ist eine gute Voraussetzung dafür, dass es sich leicht an andere Sprachen anpassen lässt. Spezielle Probleme des Deutschen, z.B. Partikelverben und Komposita, sollen, soweit es sich vermeiden lässt, in dieser Arbeit nicht behandelt werden.

Eine wichtige Grundlage für das Verfahren ist das zweisprachige Referenzmaterial, das es erlaubt, einzelne Sätze mit ihrer Übersetzung abzurufen. Dessen Aufbereitung für die Nutzung in der zum Ziel gesetzten Anwendungsperspektive wird einen großen Teil dieser Arbeit einnehmen.

Zusammengefasst ist also das Ziel meiner Arbeit, einen Ansatz zur Generierung von Übersetzungsvorschlägen auf Basis eines bilingualen Korpus soweit zu beschreiben, dass seine Realisierbarkeit beurteilt werden kann. Die Konkretisierung soll soweit gehen, dass der Ansatz zumindest manuell auf einen Testkorpus angewendet werden kann. Dabei ist klar, dass keine Ergebnis genannt oder gar eine Evalution der Übersetzungsleistung durchgeführt werden kann. Ziel soll es sein, die einzelnen Schritte des Verfahrens angemessen zu beschreiben und mit Korpusbelegen zu erläutern.

1.3 Überblick

Die Beschreibung eines Translation Memory als eine Software-Komponente, die das Referenzmaterial nach dem zu übersetzenden Satz durchsucht und automatisch die dort vorliegende Übersetzung für die aktuelle Übersetzung übernimmt, ist für das Verständnis der Funktionsweise und der Probleme, die sich dem Anwender oder dem Entwickler eines TM-Systems stellen, unzureichend. Kapitel 2 geht daher auf die Grundlagen ein. Es beschreibt, wie ein TM in den Übersetzungprozß eingebunden ist und wie es funktioniert, insbesondere wie es die Übersetzung findet. Das Grundlagenkapitel endet mit einer kurzen Beschreibung der verwandten Themen „Concordancing“ und „maschinelle Übersetzung“ und grenzt sie von Translation Memory ab.

Voraussetzung für die Benutzung eines Translation Memory ist, dass bereits übersetzter Text vorliegt.[9] Um mit einer TM-Erweiterung experimentieren zu können benötigt man eine möglichst umfangreiche Sammlung von Texten zusammen mit ihrer Übersetzung, ein bilinguales Korpus, das auf Satzebene alignt ist. Mir steht das Korpus des Studienprojekt KoKS und weiteres Material aus Kummer und Wagner (2002) zur Verfügung. In Kapitel 3 werden das von mir verwendete Korpus und die Schritte beschrieben, die nötig sind, um die

[9] Zwar kann der Übersetzer mit einem leeren Translation Memory seine Arbeit beginnen. Aber erst wenn zumindest ein Satz übersetzt wurde und zusammen mit dem Ausgangssatz ins Referenzmaterial aufgenommen wurde, kann das Translation Memory in Aktion treten.

Texte für die Benutzung im Translation Memory aufzubereiten. Besonders ausführlich werde ich die Annotation der Wortarten (POS-Tagging) darstellen, da sich mein Ansatz durch die Nutzung der Wortarteninformation von einfachen TMs unterscheidet. Abgeschlossen wird das Kapitel mit der Ermittlung einer Stichprobe von Beispielsätzen, zu denen Fuzzy-Matches gesucht und klassifiziert werden.

Kapitel 4 stellt dann den Ansatz zum Kombinieren mehrerer nur teilweise übereinstimmender Fundstellen im Referenzmaterial vor. Es werden Möglichkeiten zur Umsetzung aufgezeigt, die sich auf die in den vorangehenden Kapiteln entwickelten Grundlagen stützen. Das Kapitel schließt mit einer kurzen Bewertung ab.

Kapitel 2

Translation Memory in CAT

In diesem Kapitel wird die Funktionsweise von Translation Memorys beschrieben. Zuerst wird kurz verdeutlicht, wie sie beim Übersetzen eingesetzt werden. Dann wird darauf eingegangen, wie ein Translation Memory arbeit. Zwei Phasen werden dabei unterschieden. Vor der eigentlichen Übersetzungstätigkeit wird das in zwei Sprachen vorliegende Textmaterial, der Bitext, segmentiert. In der Übersetzungsphase wird dieses aufbereitete Material benutzt, um Übersetzungsvorschläge abzurufen. Interessant ist hier der Fall, der eintritt, wenn keine exakte Übereinstimmung im Referenzmaterial gefunden werden kann. Dann wird eine ähnliche Textstelle gesucht, um doch noch eine Übersetzung automatisch erzeugen zu können. In die Beurteilung der Textstellen können Ergebnisse einer liguistischen Analyse einfliessen. Nach einer kurzen Bemerkung zur Berücksichtigung von Layout-Informationen folgt eine Zusammenstellung von Evaluationskriterien.

2.1 Integration in den Übersetzungsprozess

In diesem Abschnitt soll ein Eindruck davon vermittelt werden, wie ein TM eingesetzt werden kann. Auf andere Werkzeuge, die dem Übersetzer zur Verfügung stehen, bin ich bereits in der Einleitung kurz eingegangen. Bowker (2002) liefert eine allgemeinverständliche Einführung in CAT.

2.1.1 Anwendungsszenario

Ein Translation Memory kann seinen vollen Nutzen nur entfalten, wenn bereits übersetzte Texte zur Verfügung stehen, die große Ähnlichkeit mit dem zu übersetzenden Text haben. Der erste Schritt ist daher, solche Texte in das Translation Memory einzuspielen. Wenn das Textmaterial dabei auf viele Dokumente[1] verteilt ist, stellt sich die Aufgabe, die ausgangssprachlichen Texte und die Übersetzungen einander zuzuordnen. Eventuell können hier Regeln, nach denen die Dokumente benannt wurden, ausgenutzt werden. Z.B. werden häufig die Zusätze `de` und `en` verwendet, um deutsche und englische Texte zu unterscheiden. Sofern der Name der Dokumente nicht mit übersetzt wurde, gibt dann der restliche Teil der Dateinamen die Zuordnung vor. Mir ist nicht bekannt, ob TM-Produkte verfügbar

[1] Dokument und Datei verwende ich synonym. Anders als Begriffe wie Absatz, Abschnitt und Kapitel, die sich auf die Gliederung eines Textes beziehen, bezeichnet Dokument die größte zusammenhängend gespeicherte Texteinheit.

Zuordnung	Englisch	Deutsch
1	Germany indicates acceptance of Zaire 's new leadership	Bundesregierung signalisiert den neuen Machthabern in Zaire Akzeptanz
2	According to the German Foreign Ministry the German government is basically ready to collaborate with the new leaders of Zaire.	Die Bundesregierung ist nach Angaben des Auswärtigen Amtes grundsätzlich bereit, mit den neuen Machthabern in Zaire zusammenzuarbeiten.
3	A Foreign Ministry spokesperson said the question of recognizing the government was not an issue since Germany recognizes states but not governments.	Ein Sprecher des Auswärtigen Amtes in Bonn sagte, die Frage der Anerkennung der neuen Regierung stelle sich nicht. Die Bundesregierung erkenne lediglich Staaten, nicht aber Regierungen an.
4	The spokesperson called for immediate, free, and fair elections in Zaire.	Der Aussenamtssprecher forderte rasche, freie und faire Wahlen in Zaire.

Abbildung 2.1: Satzentsprechungen im Dokumentenpaar

sind, die die Zuordnung automatisch anhand von Merkmalen der Dokumente wie z.B. Anzahl und Länge der enthaltenen Absätze durchführen.

Im nächsten Schritt wird die Zuordnung gewöhnlich bis auf die Satzebene verfeinert, um während der Übersetzungsarbeit die Übersetzung eines Satzes abrufen zu können. Dazu müssen Satzgrenzen erkannt und zwischen den Sprachseiten einander zugeordnet werden. Welche Probleme dabei auftreten können und warum eine Automatisierung schwierig ist, wird in den Abschnitten 2.2 und 3.2 erläutert. In der Regel muss der Benutzer unterstützend eingreifen. Abbildung 2.1 zeigt einen Teil des Dokumentpaars de-news/1997-05-19.de.asc - de-news/1997-05-19.en.asc, das entsprechend aufbereitet wurde.

Nach der Vorbereitung des Referenzmaterials kann mit der Erstellung der Übersetzung begonnen werden. Wie nun der weitere Ablauf aussieht, insbesondere wann das Translation Memory Übersetzungen vorschlägt, hängt von der CAT-Software ab, die versucht, die einzelnen Übersetzungswerkzeuge mit einem speziellen Bedienungskonzept besonders gut miteinander zu verbinden und in den Arbeitsablauf zu integrieren. Im Folgenden wird beispielhaft ein Ablauf beschrieben, bei dem das Translation Memory nicht integriert sondern als eigenständige Komponente verwendet wird.

Denkbar wäre z.B., dass die Sätze des zu übersetzenden Ausgangstextes durchnummeriert werden und ein zweites Dokument für die Übersetzung erstellt wird, das die gleichen Nummern enthält. Das Translation Memory trägt fertige Übersetzungen für die Sätze ein, die es im Referenzmaterial findet. Die weitere Texteingabe könnte dann in einem normalen Textverarbeitungssystem erfolgen. Der Übersetzer kann auf den nummerierten Ausgangstext zugreifen und anhand der Nummern erkennen, welche Bereiche bereits durch das

Translation Memory übersetzt wurden. Wenn der zu übersetzende Text viele Wiederholungen von Passagen oder Einzelsätzen enthält, die nicht im Referenzmaterial vorhanden sind, dann kann der Übersetzer in diesem Beispielszenario davon profitieren, von Zeit zu Zeit die neuen Satzpaare in das Translation Memory einzuspielen und mit einer neuen Zieltextvorlage weiterzuarbeiten. Da die Sätze nummeriert sind, entfällt das aufwendige Zuordnen der Übersetzungen, das beim Einspielen neuer Texte ins Translation Memory notwendig ist.

Das Beispiel macht deutlich, dass isolierte CAT-Werkzeuge und der Rückgriff auf Standardwerkzeuge, die für die monolinguale Textverfassung gedacht sind, keine optimale Arbeitsumgebung für den Übersetzer sind. Im Beispiel wurde die Verbindung zwischen Ausgangstext und Übersetzung durch die Nummerierung der Sätze hergestellt. Wünschenswert wäre jedoch, dass das Abrufen der sich entsprechenden Textstellen von der CAT-Software erleichtert wird. Besser in den Übersetzungsprozess integrierte Translation Memorys sind des Weiteren inkrementell, d. h. sie übernehmen automatisch fertig übersetzte Sätze mit ihrer Übersetzung ins Referenzmaterial, sodass sie sofort für andere, noch zu übersetzende Sätze zur Verfügung stehen.

2.1.2 Einsatz für mehrere Übersetzungsaufträge

Das vom Übersetzer erstellte Textmaterial ist zusammen mit den Ausgangstexten eine wertvolle Informationsquelle für spätere Übersetzungen. Wie im vorangegangenen Abschnitt beschrieben, kann das Material mit einem Translation Memory genutzt werden. Darüber hinaus ermöglichen Concordancer und Terminologie-Extraktion einen tieferen Einblick in die getroffenen Übersetzungsentscheidungen. Diese wertvolle Ressource möchte ein Übersetzer natürlich auch für spätere Übersetzungsaufträge nutzen.

Das Textmaterial sollte mindestens nach Auftraggeber und Textsorte geordnet archiviert werden, damit bei späteren Aufträgen das zu verwendende Material passend eingeschränkt werden kann. Z.B. kann die verwendete Terminologie in der Ausgangs- und/oder Zielsprache zwischen einzelnen Auftraggebern so sehr abweichen, dass die Verwendung keinen Nutzen bringt. Die Textsortenbeschränkung kann Sinn machen, um nicht beim Korrekturlesen darauf achten zu müssen, ob das Translation Memory womöglich Material eingesetzt hat, dessen sprachliche Merkmale von den Textkonventionen des zu übersetzenden Textes abweichen, oder um die Suche nach relevanten Informationen dadurch zu beschleunigen, dass Texte, von denen keine Suchergebnisse erwartet werden, von der Suche ausgeschlossen werden.

Der Nutzung der Texte stehen aber häufig die Wünsche des Auftraggebers im Wege. Bei Dokumenten, die nicht für die Öffentlichkeit bestimmt sind, kann leicht nachvollzogen werden, warum sie nach der Fertigstellung des Auftrags nicht beim Übersetzer verbleiben sollen. Die Richtlinien für den Umgang mit betriebsinternen Dokumenten werden oft reflexartig auch auf solche Dokumente angewendet, die öffentlich zugänglich sind, wie z.B. Bedienungsanleitungen. Vielleicht spielt aber auch die Angst davor eine Rolle, dass ein Fremder, womöglich ein Konkurrent, von der bezahlten Leistung profitieren könnte.

Andererseits kann der Auftraggeber dem Übersetzer selbst Textmaterial zur Verfügung stellen. Im Falle der Verwendung mit einem Translation Memory bedeudet dies in der Regel, dass die Zuordnung der Sätze von Ausgangs- und Zieltext erneut aufgebaut werden muss, da nur die Texte nicht jedoch das bei der Übersetzung benutzte TM vorhanden sind oder da gar kein Translation Memory eingesetzt wurde. Aus ökonomischen Gründen muss der Übersetzer bei sehr großen Textbeständen eine Auswahl von Texten treffen, die ins Translation Memory eingespielt werden. Anders sieht die Situation aus, wenn beim

Auftraggeber zusätzlich zu den Dokumenten auch eine Repräsentation der Zuordnung vorhanden ist.

2.1.3 Austauschbarkeit mittels TMX

Ein Translation Memory speichert die Zuordnung der Sätze (oder anderer Einheiten) von Ausgangs- und Zieltext dauerhaft zusammen mit den Texten. Die einmal geleistete Arbeit, diese Zuordnung herzustellen, kann somit in beliebig vielen Übersetzungsprojekten genutzt werden. Allerdings ist die Art, wie diese Daten gespeichert werden, nicht einheitlich. Jeder Hersteller löst auf eigene Weise das Problem, die Daten so zu speichern, dass die gewünschten Zugriffsarten ausreichend schnell erfolgen können. Auch ist unterschiedlich, welche zusätzlichen Informationen über das Referenzmaterial abgelegt werden. Die Daten eines TMs können daher nicht direkt in einer TM-Software eines anderen Herstellers benutzt werden.

Bis vor kurzem gab es auch keine Möglichkeit, die Daten in einem von mehreren Herstellern unterstützten Austauschformat zu exportieren oder zu importieren. Wenn Referenzmaterial in einem anderen TM genutzt werden sollte, musste das Textmaterial erneut sämtliche Vorverarbeitungsschritte einschließlich der aufwendigen Zuordnung der Übersetzungsentsprechungen durchlaufen. Dies bedeutete zum einem, dass dem einzelnen Übersetzer eine Hürde gestellt wurde, das Produkt zu wechseln. Zum anderen musste der Übersetzer verschiedene TM-Software besitzen oder zumindest mit ihnen umgehen können, um vorhandene Translation Memorys nutzen zu können, die der vorausschauende Auftraggeber bereits zusammen mit den bisher erstellten Übersetzungen erworben hat. Des Weiteren werden die Verwendungsmöglichkeiten des vorhandenen Referenzmaterials eingeschränkt, wie Alesiani (1997) darlegt. Alesiani betrachtet die Formate, in denen die zu übersetzenden Texte und das Referenzmaterial verarbeitet werden können. Er behauptet, dass das Übersetzungswissen des Translation Memorys kaum genutzt werden könne, wenn das zu übersetzende Textmaterial in einem anderen Format vorliegt als das, in dem die Dokumente kodiert waren, die dem vorhandenen Referenzmaterial zus Grunde liegen. Diese Aussage macht nach der bisherigen Beschreibung von TM in dieser Arbeit keinen Sinn. Ein Übersetzungsvorschlag wird von TM immer dann unterbreitet, wenn ein zu übersetzender Satz im Referenzmaterial gefunden wird. Man würde annehmen, dass nur die Wörter oder Zeichen des Satzes aber nicht das Format für die Suche eine Rolle spielen. Eine Erklärung der Behauptung könnte sein, dass Alesiani TM-Produkte verwendet, die formatspezifische Informationen mit im TM speichern und bei der Suche diese auch berücksichtigen. Wie dem auch sein, zu der Schlussfolgerung, dass Barrieren bei der Nutzung des in einem TM gespeicherten Übersetzungswissens bestehen, gelangt man auch einfacher. Kein TM unterstützt alle denkbaren Formate. Liegen in einem Übersetzungsauftrag inhaltsähnliche Dokumente in verschiedenen Formaten vor, dann kann es passieren, dass dem Übersetzer kein TM-Produkte zur Verfügung steht, dass beide Formate unterstützt. Es müssen zwei verschiedene TM-Produkte eingesetzt werden. Die zu den Dokumenten im einen Format gemachten Übersetzungen können so nicht bei der Übersetzung der Dokumente im anderen Format genutzt werden. Die drei Formatgebiete, die man in der Softwarelokalisierung antrifft — nämlich Benutzerschnittstellen, Hilfesysteme und Handbücher.[2] —, bilden laut Alesiani solche Inseln, die nicht überbrückt werden können. Er nennt dies Fragmentierung.

Zur Lösung dieser Probleme hat die Localization Industry Standards Association (LISA) im September 1997 ein standardisiertes Format für den Austausch von Translation

[2] Genannt werden die Begriffe „Software“, „Help“ und „Publications“. Die Übersetzung trifft vermutlich das gemeinte, da später von „WinHelp“ und „software resource files“ die Rede ist.

Memorys vorgestellt. Es gehört zu einer Gruppe von Standards (OSCAR, Open Standards for Container/Content Allowing Re-use) und heißt Translation Memory eXchange (TMX). Es gehört zu einer Gruppe von Standards, die unter der Bezeichnung OpenTag zusammengefasst werden. Weitere OpenTag Standards sind TBX für den Terminologieaustausch und OLIF für Wörterbücher. Drei Hersteller durchlaufen zur Zeit den Zertifizierungsprozess für die Konformität zu diesem offenen, herstellerübergreifenden Austauschformat.

Laut Erjavec (1999) (Artikel zu einem ELAN Korpus) wird in Melby (1998) TMX vorgestellt. TMX beschäftige sich vor allem damit, Markup-Informationen der Ausgangsdokumente möglichst gut zu erhalten. Andererseits sagt Erjavec, dass das TMX-Format Strukturinformation (er nennt das DIV-Tag) nicht repräsentiert würde. Des Weiteren erwähnt er dort das Text Encoding and Interschange Format TEI P3 als Alternative. TEI erhalte die ursprünglichen Dokumente und speichere die Zuordnung in einer getrennten Datei.

Ein weiteres offenes Format, das für den Austausch von TMs adaptiert werden könnte, ist CES (Corpus Encoding Standard).

2.2 Segmentierung eines Bitexts in TUs

Ein Translation Memory wird eingesetzt, um Textstücke, die bereits einmal übersetzt wurden, nicht von neuem übersetzen zu müssen, sondern die Übersetzung vom Computer automatisch einfügen zu lassen. Wichtig für ein TM ist es, dass es die Übersetzung der vorgegebenen Einheit im zweisprachigen Textmaterial zuverlässig identifizieren kann. Der Ausgangs- und Zieltext des Referenzmaterials werden dazu so segmentiert, dass das i-te Segment des Zieltextes die Übersetzung des i-ten Segments der Ausgangstextes ist. Innerhalb der Segmentpaare können andere Zuordnungen als 1 : 1 Zuordnung vorliegen, wie in diesem Abschnitt noch erläutert wird. Insbesondere ist es möglich, dass ein Segment mehrere Einheiten umfasst oder leer ist. Sind die Einheiten Sätze, dann wird häufig genau eine Einheit der Ausgangssprache einer Einheit der Zielsprache zugeordnet. Ein Segmentpaar, dessen zielsprachlicher Teil Übersetzung des ausgangssprachlichen Teils ist, wird als Translation Unit (TU) bezeichnet. Segmentpaare von verschiedenen Segmenten mit gleichem Index sind also nach Wahl der Segmentierung immer TUs.

Die Zuordnung der Einheiten, die z. B. durch die im vorangegangenen Absatz beschriebene Segmentierung vorgegeben werden kann, wird Alignment genannt. Am Ende dieses Abschnitts wird in Grundzügen auf die verschiedenen Arten von Alignments ein eingegangen, die zum Teil flexibler als die beschriebene Segmentierung sind. Zunächst jedoch wird ein kurzer Überblick gegeben, welche Einheiten in Frage kommen und welche Probleme sich mit ihnen verbinden. Dann werden speziell Sätze als Einheiten betrachtet. Es wird mit Material aus dem in dieser Arbeit verwendeteten Korpus verdeutlicht, dass selbst bei Sätzen das Auftreten von nicht trivialen Übersetzungsentsprechungen nicht ungewöhnlich ist.

Konkrete Verfahren, mit denen ein Alignment hergestellt werden kann, werden in diesem Abschnitt nicht beschrieben. Hier soll nur ein grundlegendes Verständnis geschaffen werden. Ein Beispiel für ein Verfahren kann im Abschnitt 3.2.5 gefunden werden, in dem der für diese Arbeit verwendete maschinelle Satzaligner vorgestellt wurd. Eine Übersicht über Alignmentverfahren bietet Véronis (2000). Viele CAT-Programme alignen jedoch nicht automatisch, sondern unterstützen den Benutzer lediglich bei der Herstellung eines Alignments.

2.2.1 Granularität

Die Art der sprachlichen Einheit hat wesentlichen Einfluss darauf, wie leicht die Identifikation der Übersetzung fällt. Werden z.B. nur vollständige Absätze im TM nachgeschlagen, dann kann die Übersetzung zu einer Fundstelle leicht im zielsprachlichen Text erkannt werden, da die Absatzgrenzen in elektronischen Dokumenten eindeutig sind[3] und gewöhnlich Absätze eins zu eins übersetzt werden, d.h. Absätze der Ausgangstextes werden nicht aufgeteilt oder vereinigt und auch nicht umgeordnet. Wenn z.B. die Fundstelle der elfte Absatz des Ausgangstextes ist, dann kann das Translation Memory als Übersetzungsvorschlag einfach den elften Absatz des Zieltextes ausgeben.

Schwieriger wird es, wenn nicht von einer 1 : 1 Übersetzung der Einheiten ausgegangen werden kann. Bereits Sätze verletzen diese Bedingung so häufig und unregelmäßig, dass die Satzpositionen nicht einfach umgerechnet werden können, wie im nachfolgenden Abschnitt demonstriert wird. Bei kleineren Einheiten wird es zunehmend schwieriger, die Grenzen der Einheiten auf beiden Sprachseiten zuverlässig zu erkennen. Eine 1 : 1 Zuordnung der Einheiten ist häufig nicht möglich.

Mit der Größe der Einheiten ist dabei nicht die Länge in Wörtern oder Zeichen gemeint, sondern die (syntaktische) Kategorie, zu der die Einheit gehört, die im TM nachgeschlagen werden soll. Genauer gesagt ist die Position innerhalb der Hierarchie der Kategorien relevant. Hier werden in der Syntax gewöhnlich lexikalische Kategorien (Nomen, Verb etc.), phrasale Kategorien (z.B. Nominalphrasen), Teilsätze und Sätze unterschieden. Absätze erweitern diese Hierarchie nach oben hin, indem sie eine Folge von Sätzen umfassen. In der anderen Richtung wäre denkbar, die Gliederung der Lexeme in Morpheme hinzuzunehmen. Letzteres könnte auf ein Translation Memory hinauslaufen, das die Bestandteile von Komposita, die Ausgangsformen von Derivationen und Stamm und Flexionsendung flektierter Formen als Einheiten behandelt und entsprechend eine Übersetzung unabhängig vom Kontext aus dem Referenzmaterial liefert. Hier spricht man aber gewöhnlich nicht von Translation Memorys. Systeme, die einzelne Wörter als Einheiten wählen, würde man eher als eine sehr einfache Form der Lexikonextraktion bezeichnen.

Jede dieser Größenstufen erfordert angepasste Methoden, um die Einheiten in der Übersetzung erkennen und der Suchvorgabe zuordnen zu können. Die Art der Behandlung von Absätzen wurde weiter oben schon angedeutet. Bei Sätzen gibt es verschiedene Verfahren, die sich u.a. darin unterscheiden, wie weit sie den Text analysieren und welches linguistische Wissen sie investieren. Die meisten Schwierigkeiten bereitet es, die Zuordnungen richtig zu erkennen, die nicht 1 : 1 verlaufen. (Siehe unten.) In der Praxis überlassen die Translation Memorys der CAT-Anbieter es dem Benutzer, einmal zu Beginn der Arbeit die Zuordnungen für das gesamte Textmaterial herzustellen.

Sind die Einheiten kleiner, dann tritt das Problem der Erkennung der Einheiten in den Vordergrund. Dieses Problem wird gewöhnlich mit computerlinguistischen Methoden wie Parsing oder Chunking gelöst. Es wird also linguistisches Wissen über die Wortarten der Wörter und ihre grammatische Struktur benötigt. Dies bedeutet zum einen, dass der Anpassungsaufwand der Verfahren an weitere Sprachen, die der TM-Softwareanbieter mit seinem Produkt unterstützen will, sehr groß ist und dass häufiger falsche Einheiten identifiziert werden, da nicht alle Ausnahmen und Sonderfälle natürlicher Sprachen berücksichtigt wer-

[3]Dies trifft nicht immer zu. Im Format „Nur Text“ gibt es zwar die Konvention, entweder jeden Absatz als eine lange Zeile zu repräsentieren und es dem Textverarbeitungsprogramm zu überlassen, bei der Bildschirmausgabe oder beim Drucken Zeilenumbrüche einzufügen, oder aber nach jedem Absatzende eine Leerzeile zu speichern. Diese Konvention wird aber mitunter verletzt, oder es wird bei der Verarbeitung der Dokumente nicht beachtet, welche Bedeutung die Zeilenumbrüche haben. Des Weiteren kann bei der optischen Erfassung (OCR) gedruckter Dokumente die korrekte Erkennung der Absatzgrenzen fehlschlagen, wenn die Vorlage keine Anhaltspunkte (z.B. Einrückungen) enthält, die die OCR-Software nutzen kann.

den können. Zum anderen wird das Gebiet sowohl theoretisch als auch praktisch seit Jahren untersucht, sodass ein TM-Entwickler auf bestehende Softwarekomponenten zurückgreifen kann. Die Zuordnung der Einheiten der Übersetzung zu den Einheiten des Ausgangstextes gestaltet sich auf der Ebene der Phrasen aber nicht einfach, da die Reihenfolge verändert sein kann und Verschmelzungen und Aufteilungen wesentlich häufiger sind als bei Sätzen.

Wiederverwendbarkeit

Ein anderer Aspekt, der durch die Größe der Einheit beeiflusst wird, ist die Wiederverwendbarkeit der Übersetzung in anderen Kontexten. Es reicht nicht aus, wenn der Übersetzungsvorschlag des Translation Memorys eine belegte Übersetzung der Ausgangseinheit ist. Er muss auch in den neuen Kontext passen. Handelt es sich bei der Einheit des Translation Memorys um einen Satz, dann sind meistens in der Einheit selbst genügend Kontextinformation vorhanden, die auftretende Wörter aus offenen Wortklassen wie z. B. Verben und Nomen disambiguieren. Häufig bereiten aber Anaphern Probleme, z. B. Pronomina, wenn ihr Bezugspunkt außerhalb des Satzes liegt.

(2.1) The drive has a 2 MB buffer, [...]. It spins at 5400 RPM. (John Martin, Southern Nevada User's Group)

(2.2) The fan is the problem. [...] It spins at 5000 rpm. (www.pureoc.com)

Im Beispiel (2.1) geht es um eine Computerfestplatte. Eine naheliegende Übersetzung des Antezendens „the drive" ins Deutsche ist Femininum (Festplatte) oder Neutrum (Gerät) nicht aber Maskulinum, wie es im Beispiel (2.2) der Fall ist, da es hier um einen Lüfter zur Kühlung eines Computers geht und somit die Alternativen „Gebläse" oder „Triebwerk" keine geeigneten Übersetzungen sind.[4]

Das Genus muss aber nicht verschieden sein, damit die Übersetzung nicht in den Kontext passt. Zum einen übernimmt ein Übersetzer das Muster von Nomen und Pronomen i. d. R. nichts aus dem Ausgangstext, sondern entscheidet bei der Produktion des Zieltextes neu, wann eine Wiederholung des Nomens und wann ein Pronomen die Lesbarkeit erhöht. Wenn im Beispiel (2.1) der Übersetzer nochmal das Nomen „Festplatte" aufgreift, dann würde das Translation Memory dem Übersetzer zu (2.2) einen Übersetzungvorschlag unterbreiten, der inhaltlich falsch ist und den Übersetzer irritiert.

Zum anderen kommt es vor, dass durch das veränderte Antezedens weitere Wörter des Satzes betroffen sind. Im obigen Beispiel würde dies eintreten, wenn der Text

(2.3) The washing machine provides best spin performance. It spins at 1200 RPM. (konstruiertes Beispiel)

zu übersetzen wäre, nachdem die ersten beiden Beispieltexte übersetzt und im Translation Memory gespeichert wurden. Zwar ist in einem der beiden Übersetzungsvorschläge, die das Translation Memory macht, das Pronomen Femininum. Aber die Übersetzungen des Verbs „spin", die aus anderen Kontexten stammen, sind hier zu unspezifisch. Das Verb sollte mit „schleudern" übersetzt werden und nicht mit „rotieren", „laufen" oder „drehen".

Diese Unsicherheiten können immer dann auftreten, wenn die zu übersetzende Einheit nicht genug Informationen enthält. Man könnte also vermuten, dass der Anteil der geeigneten Übersetzungsvorschläge des Translation Memorys mit der Länge der Einheit

[4]Im Beispiel sind die Zahlen und Einheiten nicht identisch. Ein striktes Translation Memory würde daher keinen Übersetzungsvorschlag liefern, wenn nur (2.1) mit einer Übersetzung in der Datenbasis stünde und der letzte Satz von (2.2) zu übersetzen wäre. Praxistaugliche Translation Memorys erkennen jedoch Zahlen und Einheiten und zeigen auch Fundstellen an, die kleine Änderungen enthalten.

Korpus	Deutsch	Englisch	Verhältnis	Ausgangssprache
DE-News	64 432	59 732	1,08	Deutsch
EU	112 828	110 326	1,02	unbekannt
Harry Potter	35 902	36 951	0,97	Englisch
Gesamt	213 162	207 009	1,03	—

Tabelle 2.1: Anzahl der Sätze in den verwendeten Korpora

in Wörtern oder Zeichen zunimmt. Allerdings ist die Länge der Einheit nur ein Anhaltspunkt. Auch eine sehr lange Einheit kann eine Abhängigkeit vom Kontext enthalten, die die Verwendung der Übersetzung in einem anderen Kontext verhindert.

Bei kleineren Einheiten als dem Satz stehen der Wiederverwendbarkeit der im Translation Memory gespeicherten Übersetzungen die in den vorangehenden Absätzen beschriebenen Probleme wesentlich häufiger im Weg. Wenn Subjekt, Verb, Objekte und Adjunkte isoliert abgerufen werden, fehlt der Kontext ganz, der Mehrdeutigkeiten auflösen könnte. Das ist auch ein Grund, warum man sich meistens auf Sätze als Einheiten beschränkt.

2.2.2 $n : m$ Übersetzung von Sätzen

Der Leser, der eine Fremdsprache gelernt hat, wird sicherlich nachvollziehen können, dass auf der Ebene der Wörter und Phrasen keine sequentielle 1 : 1 Zuordnung möglich ist, d. h. dass die Reihenfolge dort nicht immer beibehalten, ein Wort durch mehrere wiedergegeben oder auch Teile kompakter übersetzt werden können. Aber bei Sätzen und Absätzen liegt dies nicht gleich auf der Hand. In diesem Unterabschnitt werden daher Zahlen aus dem Korpus vorgestellt, das in dieser Arbeit verwendet und in Abschnitt 3.1 ausführlich vorgestellt wird.

Wenn man die Anzahl der Sätze im Ausgangstext mit der der Übersetzung vergleicht, dann ergibt sich eine untere Schranke für die Anzahl der Sätze, die nicht 1 : 1 einem anderen Satz zugeordnet werden können. Man könnte vermuten, dass beim Übersetzen bestimmte Alignment-Beads (neben den dominierenden 1 : 1 Zuordnungen) unabhängig vom Übersetzer oder von der Übersetzungsrichtung bevorzugt werden. Tabelle 2.1 zeigt ein etwas unklareres Bild. Der Rückgang der Anzahl der Sätze um über sieben Prozent bei den DE-News Texten kann darauf zurückzuführen sein, dass Freiwillige die Übersetzungen anfertigen, die zum größten Teil vermutlich nicht dazu ausgebildet sind. Des Weiteren werden einzelne Nachrichten gar nicht übersetzt, wenn die Redaktion der Ansicht ist, dass sie für die Leser uninteressant sind. Auch wurde beobachtet, dass zum Ende einer Nachricht weitere Ausführungen des Ausgangstextes ausgelassen wurden.

Bei den Harry Potter Büchern dürften andere Gründe vorliegen. Sie wurden von Bewunderern der Bücher eingescannt und am Computer in Text umgewandelt. In einer kleinen Stichprobe fällt auf, dass einige offensichtliche Kommata von der OCR als Punkte interpretiert wurden. Des Weiteren hat der Übersetzer an Stelle eines Semikolons gelegentlich einen Punkt gesetzt.[5] Das Korpus EU[6] verhält sich dagegen wie erwartet, wenn man

[5] Merkel (2001) beobachtet ähnliche Unterschiede in der Verwendung dieser Satzzeichen in einer schwedischen Übersetzung eines englischen Romans. Nach Anpassung der Satzgrenzenerkennung erhöht sich der Anteil der 1 : 1 Zuordnungen auf das Niveau der anderen von Merkel alignten Texte. (Seite 3–4).

[6] Die verwendeten Kurzbezeichnungen für die Korpora werden im Abschnitt 3.1 eingeführt.

Korpus	0 : 1 / 1 : 0	1 : 1	2 : 1	1 : 2	andere
DE-News	1	51 027	3 951	1 151	1 469
EU	1	88 511	5 896	4 417	3 003
Harry Potter	0	29 008	1 477	1 898	994
Gesamt	2	168 546	11 324	7 466	5 466

Tabelle 2.2: Satzzuordnungen in den verwendeten Korpora

Korpus	0 : 1 / 1 : 0	1 : 1	2 : 1	1 : 2	andere
DE-News	0,0 %	88,6 %	6,9 %	2,0 %	2,6 %
EU	0,0 %	86,9 %	5,8 %	4,3 %	2,9 %
Harry Potter	0,0 %	86,9 %	4,4 %	5,7 %	3,0 %
Gesamt	0,0 %	87,4 %	5,9 %	3,9 %	2,8 %

Tabelle 2.3: Anteil der Satzzuordnungen

annimmt, dass die Texte mehrheitlich von Englisch nach Deutsch übersetzt wurden. Die Zunahme der Anzahl der Sätze durch das Übersetzen um etwa zwei bis drei Prozent kann durch die Neigung von Übersetzern, Sätze aufzuteilen, erklärt werden.

Um ein genaueres Bild zu erhalten, muss man untersuchen, welche Zuordnungen tatsächlich auftreten. Einen guten Anhaltspunkt gibt das maschinell erstellte Alignment. Die Häufigkeiten der verschiedenen Zuordnungen sagen mehr aus als die Satzanzahlen. Wenn z. B. neben 1 : 1 Zuordnungen nur vereinzelt $m : 0$ Zuordnungen mit großem m aufträten, dann würde dies bedeuten, dass lediglich einige Textpassagen weggelassen wurden. Hingegen würden viele 1 : 2 Zuordnungen die Annahme stützen, dass Übersetzer zum Aufteilen von Sätzen neigen.

Tabelle 2.2 zeigt die Häufigkeiten der wichtigsten Zuordnungsarten aufgeschlüsselt nach den verwendeten Korpora. Unabhängig von der Übersetzungsrichtung bedeutet hier eine $n : m$ Zuordnung, dass n Sätze des Deutschen zusammen mit m Sätzen des Englischen eine Translation Unit bilden. Die Richtung der Übersetzung kann der Tabelle 2.1 entnommen werden. Im EU Korpus können auch Dokumente enthalten sein, die aus einer dritten Sprache übersetzt wurden. Wie groß deren Anteil am Korpus ist, kann der KoKS Dokumentation nicht entnommen werden.

Die Zahlen machen deutlich, dass wesentlich mehr von 1 : 1 Zuordnungen abweichende Zuordnungen auftreten, als notwendig wären, um die Differenz in den Satzanzahlen zu überbrücken. Beispielweise hätten beim EU Korpus 2 502 2 : 1 Zuordnungen und 107 824 1 : 1 Zuordnungen ausgereicht, um ein Alignment zwischen den Texten herzustellen. (Dabei würden durch Dokument- und Absatzgrenzen implizierte Ankerpunkte des Alignment höchstwahrscheinlich verletzt.) Tatsächlich treten mehr als fünf mal so viele, nämlich 13 317, nicht 1 : 1 Zuordnungen auf, darunter viele 1 : 2 Zuordnungen.

Vergleicht man den Anteil der 1 : 1 Zuordnungen, der in Tabelle 2.3 prozentual angegeben ist, mit den Werten aus einer Untersuchung von Merkel (2001) und den Zahlen zum ARCADE-Korpus (Véronis (2000) Seite 374–375), fällt auf, dass die Werte im KoKS-Korpus kaum variieren. Merkel findet deulich mehr 1 : 1 Zuordnungen. Nur einer von zwei Romanen kommt beim ihm mit 91 % in die Nähe des Wertes für die Harry-Potter-Bücher.

Im ARCADE-Korpus weist das literarische Teilkorpus dagegen nur zu 76 % 1 : 1 Zuordnungen auf. Das Verhältnis von 2 : 1 und 1 : 2 Zuordnungen im KoKS-Korpus scheint ungewöhnlich zu sein. Gerade bei Romanen findet Merkel nur sehr wenige 2 : 1 Zuordnungen.

Bei der Interpretation der Zahlen muss man beachten, dass unterschiedliche Aligner verwendet wurden. Der KoKS-Aligner erzeugt nur $0 : n$ oder $n : 0$ Zuordnungen, wenn in einer Sprachseite ein leerer Absatz vorliegt. Der von Merkel (2001) eingesetzte Aligner aus der Werkzeugsammlung „DAVE“ kann solche Zuordnungen in großer Zahl hervorbringen, wenn ein Text frei übersetzt ist. Zumindest folgert er im ersten Absatz vom Abschnitt 3.2 umgekehrt, dass ein Teilkorpus frei übersetzt sein müsse, da sein Aligner Löschungen und Einfügungen erkannt hat:

> The OS2 text has a strikingly high proportion of deletions (1-0) and insertions (0-1) which indicate that the translation is not particularly close to the original, but is rather a kind of communicative, more target-oriented translation.

2.2.3 Alignment

Ein Alignment[7] ist eine Zuordnungen der Einheiten von Ausgangs- und Zieltext. Jede Einheit gehört genau einer Translation Unit an. Eine Translation Unit kann sich jedoch aus beliebig vielen Einheiten der beiden Sprachseiten zusammen setzen.

Gewöhnlich wird eine andere Terminologie verwendet. Alignment ist auch in anderen Bereichen als Translation Memory wichtig. In der Fußnote 7 wird die Bioinformatik erwähnt. Die Einheiten der zu alignenden Texte werden Alignment Beads zugeordnet, die hier Translation Units sind. Im Allgemeinen müssen Alignment-Beads aber keine Translation Units sein. Zum einen ist die Anzahl der Texte nicht auf zwei beschränkt, und keiner der Texte ist als Ausgangstext ausgezeichnet. Zum anderen müssen sie auch nicht in verschiedenen Sprachen vorliegen. Beispielweise entwickeln Ghorbel et al. (2002) Alignment-Techniken zur Behandlung verschiedener Fassungen altertümlicher Texte. Ein ähnliches Gebiet ist der Vergleich der neuen Evangelien, bei denen Auslassungen und Überkreuzungen auftreten.

Die Einheiten eines Textes, die dem selben Alignment-Bead angehören, bilden eine Gruppe. Eine Translation Unit setzt sich also aus einer ausgangssprachlichen Gruppe und einer zielsprachlichen Gruppe zusammen.

Wenn von Zuordnungen gesprochen wird, gibt es mehrere Möglichkeiten dafür, was gemeint ist. Obige Definition eines Alignments kann man mathematisch mit einer Funktion b beschreiben, die die Einheiten auf Alignment-Beads abbildet. Was genau ein Alignment-Bead ist, spielt dabei keine Rolle. Es stellt lediglich die Verbindung her zwischen den Einheiten der einzelnen Texte. Als Wertebereich für b sind z. B. die natürlichen Zahlen geeignet. (Die dadurch eingeführte Ordnung der Alignment-Beads kann unabhängig von den Ordnungen der Einheiten sein.) Für ein Translation Memory ist die Ordnung irrelevant, da die Translation Units unabhängig voneinander eingesetzt werden.

Abbildung 2.2 zeigt ein Alignment von Einheiten $D_1,\ldots,D_8$ zu Einheiten $E_1,\ldots,E_8$. Die Bezeichnungen D_i und E_j sollen dabei für die Sprachseiten Deutsch (D) und Englisch (E) stehen, auch wenn diesem Beispiel kein Text zugrunde liegt. (Die Indizes i und j nummerieren die Einheiten in der Reihenfolge, wie sie in den Texten auftreten.) Das Alignment enthält sechs Alignment-Beads. Die Funktion b ist mit dem Symbol „$\mapsto$“ angegeben.

[7] Aus dem Englischen „alignment“ – „Abgleich“, „Anordnung“; die in der Vermessungskunde gebräuchliche französische Schreibung „Alignement“ wurde nicht übernommen. Folglich wird auch die Verbform „alignieren“ nicht verwendet und stattdessen „alignen“ von „to align“ benutzt. In der Bioinformatik haben sich die gleichen Bezeichnungen für das Zuordnen von DNS- und Proteinsequenzen durchgesetzt.

Text 1	Text 2
$D_1 \mapsto 1$	$E_1 \mapsto 1$
$D_2 \mapsto 2$	$E_2 \mapsto 2$
$D_3 \mapsto 1$	$E_3 \mapsto 2$
$D_4 \mapsto 4$	$E_4 \mapsto 3$
$D_5 \mapsto 5$	$E_5 \mapsto 4$
$D_6 \mapsto 4$	$E_6 \mapsto 4$
$D_7 \mapsto 6$	$E_7 \mapsto 6$
$D_8 \mapsto 6$	$E_8 \mapsto 6$

Abbildung 2.2: Ein Alignment mit sechs Alignment-Beads

Optimales Alignment

Wenn alle Zuordnungen eines Alignments korrekt sind, dann ist das Alignment zwar zulässig aber nicht zwingend so detailiert wie gewünscht. Insbesondere reicht es nicht aus, alle Einheiten einem einzigen Alignment-Bead zuzuordnen. (Außer natürlich, wenn z. B. ein Ausgangstext so frei übersetzt wurde, dass keine feinere Zuordnung möglich ist.) Was ein erwünschtes Alignment charakterisiert, wird aus folgender Definition der Optimalität eines Alignments deutlich: Ein Alignment ist optimal, wenn es zulässig ist und kein Alignment-Bead so in zwei nicht leere Beads aufgeteilt werden kann, dass die neuen Zuordnungen immer noch korrekt sind.

Die Definition der Optimalität eines Alignments setzt die Definiton der Korrektheit der Zuordnung der Einheiten zu den Alignment-Beads voraus. Die dem gleichen Alignment-Bead zugeordneten Einheiten aller Texte, die alignt werden, sollen sich in irgendeiner Form entsprechen. Im Falle des Alignments eines Ausgangstextes mit seiner Übersetzung kann diese Entsprechung die Übersetzungsentsprechung sein. Die zielsprachlichen Einheiten eines Alignment-Beads sollen eine korrekte Übersetzung der ausgangssprachlichen Einheiten bilden. Allgemein scheint die Semantik ein geeignetes Kriterium zu sein. Dies muss aber nicht so sein. Z. B. könnte man sich eine Anwendung vorstellen, in der die Absätze von Reden, die inhaltlich nicht zusammenhängen, aber von einem Autor stammen, nur nach stilistischen Merkmalen paarweise alignt werden.

Genau genommen fehlt in obiger Definition des optimalen Alignments eine Berücksichtigung der Reihenfolge der Einheiten innerhalb der zu alignenden Texte. Wenn z. B. im Ausgangstext eine Einheit doppelt vorkommt, dann wären sie nach der Definition austauschbar. Es würde keine Rolle spielen, ob das erste oder zweite Auftreten der ersten Übersetzung zugeordnet wird.[8] Das ist aber nicht gewollt. Es sollte die Zuordnung favo-

[8]Natürlich muss eine zweite Übersetzung im Zieltext vorhanden sein, mit der die verbleibende Einheit in ein Alignment-Bead gestellt werden kann, damit ein zulässiges Alignment entstehen kann.

risiert werden, bei der die Kontexte der (über ein Alignment-Bead) einander zugeordneten Einheiten sich auch entsprechen. Die Zuordnung sollte die Reihenfolge der Einheiten möglichst erhalten, d. h. Überkreuzungen und Abweichungen von 1 : 1 Zuordnungen sollten möglichst selten auftreten.

Häufig werden die möglichen Zuordnungen noch weiter eingeschränkt. Piperidis et al. (2000) stellen eine Wortzuordnung nur zwischen Wörtern her, die in Sätzen stehen, die in einem zuvor durchgeführten Satzalignment einander zugeordnet wurden. Zuordnungen zwischen Wörtern aus verschiedenen Alignment-Beads werden dadurch ausgeschlossen. So ein hierarchisches Alignment ist durchaus typisch. Der Aligner, der in dieser Arbeit verwendet wird, führt erst ein triviales Absatzalignment durch, d. h. nur 1 : 1 Zuordnungen werden erlaubt. (Hat ein Dokument in Ausgangs- und Zielsprache nicht die gleiche Anzahl von Absätzen, dann schlägt das Alignment fehl und das Dokument kann nicht weiter verwendet werden.) Das Satzalignment wird dann innerhalb der Absätze durchgeführt. So können nur Sätze einander zugeordnet werden, die in bereits einander zugeordneten Absätzen stehen. Die zu Grunde liegende Annahme ist, dass ein Übersetzer die vorliegende Absatzstruktur respektiert und keine Inhalte in andere Absätze verschiebt.

Zwei weitere Einschränkungen der Freiheit der Zuordnung, die der verwendete Satzaligner mit anderen Alignern teilt, sind der völlige Verzicht auf Überkreuzungen und die Forderung, dass je Text nur zusammenhängende Einheiten einem Alignment-Bead zugeordnet sein dürfen. Ob die letztere Bedingung bereits durch das Überkreuzungsverbot abgedeckt ist, hängt davon ab, was man genau unter einer Überkreuzung versteht. Auf eine Definition wird hier verzichtet, da sie für die Arbeit nicht wichtig ist. Ein kritischer Spezialfall sind solche $n : 0$ und $0 : m$ Zuordnungen, die zwei Einheiten unterbrechen, die zum gleichen Alignment-Bead gehören, wie die Zuordnung 5 in Abbildung 2.2. Beide Einschränkungen zusammen lassen sich formulieren als

$$\forall i, j, k : b(e_{i,j}) > b(e_{i,k}) \rightarrow j > k,$$

wobei $e_{i,j}$ die j-te Einheit des i-ten Textes ist und b die Einheiten auf die Nummern der Alignment-Bead abbildet. In Abbildung 2.2 verstoßen Einheiten in den Alignment-Beads 1, 2, 4 und 5 gegen diese Bedingung.

Die Definition eines zulässigen Alignments muss für jede dieser Einschränkungen angepasst werden, um die Begriffe Zulässigkeit und Optimalität weiter anwenden zu können. Beim Verzicht auf Überkreuzungen dürfen beispielsweise Alignments, die sich überkreuzende Zuordnungen enthalten, nicht zulässig sein. Eine Top-Down Suche nach einem optimalen Alignment beendet dann die Unterteilung von Alignment-Beads früher. Bereiche die eigentlich eine Überkreuzungen erfordern, werden dann durch eine große Zuordnung abgedeckt. In Abbildung 2.2 würden die Beads 1 und 2 durch eine 3 : 3 Zurodnung und die Beads 4 und 5 durch eine 3 : 2 Zuordnung ersetzt werden.

Viele Satzaligner verbieten zusätzlich $n : m$ Zuordnungen mit $\max(n,m) > 2$. Dies geschieht vor allem, um die Anzahl der in Frage kommenden möglichen Zurordnungen und somit die Komplexität der Suche des optimalen Alignments zu reduzieren.

In der Praxis scheitert die Bestimmung eines optimales Alignment bereits daran, dass die Korrektheit einer Zuordnung nicht eindeutig festgestellt werden kann. Die Frage, ob zwei verschiedene Sätze das gleiche ausdrücken, kann nicht zweifelsfrei beantwortet werden. Ein maschineller Aligner kann die Korrektheit einer Zuordnung nur abschätzen. Man schwächt daher die Bedingung der Korrektheit der Zuordnungen ab, indem der Grad der Übereinstimmung der einander zugeordneten Einheiten verwendet wird, um jedes Alignment zu bewerten. Der Begriff der Zulässigkeit kann dann nicht mehr angwendet werden, bzw. jedes Alignment wird zulässig. Optimalität wird nun über die skalare Größe

definiert, mit der jedes Alignment bewertet wird. Die Bewertung soll möglichst gut sein. Da die Zahl der Alignments endlich ist, gibt es immer mindestens ein optimales Alignment.

Die Bewertung eines Alignments kann neben der Übereinstimmung des Inhalts innerhalb der Alignment-Beads auch die Art der Zuordnung und die Entfernungen der Einheiten berücksichtigen. Ein maschineller Aligner hat also die Aufgabe, ein Alignment mit optimaler Gesamtbewertung zu finden. Abschnitt 3.2.5 beschreibt einen Satzaligner, der nach diesem Prinzip arbeitet.

2.3 Berücksichtigung von ähnlichen TUs

In einer überarbeiteten Fassung eines Dokuments weisen viele Sätze nur kleine Veränderungen auf. Es werden Fehler korrigiert, die Terminologie vereinheitlicht und die Reihenfolge der Wörter der Lesbarkeit Willen verbessert. Ein Translation Memory, das nur genau übereinstimmende Fundstellen berücksichtigt, zwingt den Übersetzer, jeden auch nur geringfügig veränderten Satz erneut zu übersetzen. Ein Teil des im Translation Memory vorhandenen Übersetzungswissens kann so bei überarbeiteten Dokumenten nicht genutzt werden.

Das gleiche Problem tritt auf, wenn ein vorhandenes Dokument als Vorlage für ein neues Dokument verwendet und dabei ein Großteil der Formulierungen zwar übernommen aber leicht angepasst wird. Eingängigstes Beispiel hierfür sind Bedienungsanleitungen für Nachfolgemodelle eines Produkts. In vielen Sätzen ist nur die Produktbezeichnung ausgetauscht. Häufig ist die Produktbezeichnung in Ausgangs- und Zielsprache sogar identisch. (Dies hängt sowohl von den Sprachen als auch vom Marketing ab.) Unter diesen Bedingungen wünscht sich vermutlich jeder Übersetzer, dass das Translation Memory diese Änderungen erkennt und angepasste Übersetzungsvorschläge unterbreitet. Wie einfach dies zu realisieren ist, verdeutlicht die Behelfslösung, die entsprechenden Teile des Translation Memorys in ein Austauschformat (siehe Abschnitt 2.1.3) zu exportieren, dort die Produktbezeichnungen zu ersetzen und dann die Daten wieder zu importieren.[9]

Seltsamerweise wird von Translation Memorys nicht die Möglichkeit angeboten, jedes Auftreten der Zeichenfolge A im Anfragesatz durch eine Zeichenfolge B zu ersetzen und dann im Übersetzungsvorschlag wieder B durch A (oder B' durch A') zu ersetzen. Stattdessen wird versucht, beliebige Veränderungen zu erlauben und aus den zahlreichen Fundstellen diejenige mit den geringsten Abweichungen auszuwählen (oder die besten n oder die, die eine bestimmte Bewertungsschwelle überschreiten). Dies geschieht auf Kosten der Möglichkeit, den Übersetzungsvorschlag mit einfachen Ersetzungsregeln automatisch anpassen zu können, behandelt aber zugleich die eingangs beschriebenen Probleme mit überarbeiteten Fassungen von Dokumenten.

Ein Translation Memory, das letzteren Lösungsansatz umsetzt, muss zu dem zu übersetzenden Satz auch Stellen im Referenzmaterial finden, die nicht völlig identisch sind. Es sollen Textstellen einbezogen werden, die Ersetzungen, Einfügungen, Löschungen und Umstellungen von Wörtern aufweisen. Die ungenauen Fundstellen, die auch Fuzzy-Matches genannt werden, müssen bewertet werden, damit die Fundstellen dem Übersetzer geordnet nach Relevanz angezeigt werden können.

[9] Viele Benutzer verfügen nicht über die Fähigkeit, Lösungswege dieser Art zu Computerproblemen selbst entwickeln zu können. Unterstützung durch die Software oder zumindest durch das Benutzerhandbuch ist hier notwendig.

2.3.1 Zugriff auf das Referenzmaterial

Die einfachste Art, Fuzzy-Matches zu finden, ist, alle Sätze der ausgangssprachlichen Seite des Referenzmaterial auf ihre Relevanz hin zu prüfen. Dieses Vorgehen hat aber den Nachteil, dass die Dauer der Suche das Produkt von der Anzahl der Sätze und der Dauer der Prüfung eines einzelnen Satzes ist. Durch eine Beschleunigung der Berechnung der Relevanz wird das Problem also nur verlagert, da eine Verdoppelung des Umfangs des Referenzmaterials auch den Suchaufwand wieder verdoppelt. Es sind andere Vorgehensweisen zur Ermittlung der in Frage kommenden Stellen notwendig, die sich bei steigendem Umfang besser verhalten.

Im Falle von genau übereinstimmenden Stellen reicht zur Lösung dieses Problems ein einfacher Index aus. Ein Index listet ähnlich einem Index in einem Buch alle Stellen auf, an denen ein Suchschlüssel im Text vorkommt. Der Schlüssel, mit dem im Index nachgeschlagen wird, ist hier nur kein Einzelwort, sondern der gesamte Satz.

Zum Finden von Fuzzy-Matches kann ein solcher Satzindex nicht verwendet werden. Ein Fuzzy-Match weist gewöhnlich nur wenige Änderungen auf. Die meisten Wörter stimmen also mit dem Anfragesatz überein. Ein naheliegendes Vorgehen wäre, einen Wortindex zu erstellen, der zu jedem Wort die Sätze (oder Satznummern) auflistet, in denen das jeweilige Wort vorkommt, und dann nur einzelne Wörter des Anfragesatzes für die Suche zu verwenden. Zwar würde man auf diese Weise viele Sätze untersuchen müssen, die nur wenig, u. U. nur das Anfragewort, mit dem Anfragesatz gemeinsam haben. Die Zahl der zu prüfenden Sätze kann so aber deutlich reduziert werden, insbesondere, wenn als Anfragewörter solche Wörter des Anfragesatzes ausgewählt werden, die im Referenzmaterial selten vorkommen. Mehrere Anfragewörter sind notwendig, da auch solche Fuzzy-Matches gefunden werden sollen, die das erste Anfragewort nicht enthalten. Im Allgemeinen müssen $n+1$ Anfragen an den Wortindex gestellt werden, wenn n Änderungen erlaubt sein sollen. Die Anzahl der Sätze, die jede solche Anfrage liefert, wächst mit der Größe des Referenzmaterials: Wenn das Wort X mit der Wahrscheinlichkeit p in einem Satz auftritt, dann kann man erwarten, np Sätze prüfen zu müssen, wenn n die Anzahl der ausgangssprachlichen Sätze im Referenzmaterial ist. Die Anzahl der zu prüfenden Sätze wächst also wie im ersten Ansatz linear mit dem Umfang des Referenzmaterial.[10]

Im Abschnitt 3.2.7 wird ein Index beschrieben, der die Zahl der zu prüfenden Sätze im Vergleich zu diesem Ansatz sehr klein hält aber prinzipiell das gleiche Problem hat. Eine echte Lösung des Problems ist mir nicht bekannt. Da der benutzte Ansatz auf dem vorhandenen Textmaterial mehr als befriedigend schnell läuft, habe ich nicht nach Literatur gesucht. Eine Implementation einer Fuzzy-Match-Suche wird im Abschnitt 3.4.2 beschrieben.

Baldwin und Tanaka (2000) beschreiben auf Seite 38 ihrer Vergleichsstudie zu Ähnlichkeitsmaßen (s.u.) einige Methoden zum effizienten Zugriff auf das Referenzmaterial. Beispielsweise könnten viele Sätze bereits aufgrund ihrer Länge von der Suche ausgeschlossen werden. Wie Simard und Langlais (2001) in ihrer Einleitung schreiben, kann die Suche nach Matches auch als Information Retrieval Aufgabe gesehen werden. Umfangreiche Literatur aus einem anderen Themenbereich ist also für Translation Memory relevant.

[10] Anfangs treten noch viele neue Wörter auf, die Anzahl der Einträge im Index wächst schnell und die Länge der Einträge nimmt scheinbar nur langsam zu. Mit zunehmender Größe des Index treten nicht indizierte Wörter immer seltener auf. Man könnte meinen, dass die Länge der Einträge nun schneller wachsen müsse. Das ist aber nicht der Fall. Die Wachstumsrate für den Eintrag X ist p.

2.3.2 Ähnlichkeitsmaße

Die Relevanz eines Übersetzungsvorschlags orientiert sich daran, wie sehr der Vorschlag dem Übersetzer hilft, d. h. welchen Effizienz- und Effektivitätsvorteilt er ihm bietet. Um die Relevanz abzuschätzen stehen dem Translation Memory primär der zu übersetzende Satz, die Fundstelle und die zugeordnete Übersetzung zur Verfügung. (Sekundäre Informationsquellen sind die alternativen Fundstellen, die gesamten im TM gespeicherten Korpora und sonstige Quellen wie z. B. Wörterbücher.) Wichtigstes und naheliegendstes Kriterium ist der Grad der Übereinstimmung der Fundstelle mit dem zu übersetzenden Satz. Sie wird mit einem Ähnlichkeitsmaß gemessen. Häufig sind diese Maße symmetrisch, d. h. die Richtung des Vergleichs spielt keine Rolle.

Ob ein Wort eingefügt oder gelöscht wird, hat natürlich unterschiedlichen Einfluss auf die Nützlichkeit der Übersetzung. Vermutlich ist es für den Übersetzer einfacher, ein Wort aus dem Übersetzungsvorschlag zu entfernen als eine passende Übersetzung für ein eingefügtes Wort suchen zu müssen. Die Auswirkungen der Änderungen können aber komplexer sein, so dass auch scheinbar einfache Fälle schwierigere Anpassungen erfordern. Es wäre sinnvoll, dies experimentell zu untersuchen, um ein asymmetrisches Ähnlichkeitsmaß entwerfen zu können, das auf die spezielle Problemstellung des Translation Memorys eingeht.

Denkbar wäre auch, Kriterien in die Bewertung der Relevanz einfließen zu lassen, die die Übersetzung isoliert betrachten, wie die Komplexität der grammatischen Struktur oder den lexikalischen Schwierigkeitsgrad, den z. B. Wible et al. (2002) aus der Häufigkeit der auftretenden Wörter im Gesamtkorpus ermitteln. So könnten verständlichere Übersetzungsvorschläge bevorzugt werden, die i. d. R. auch leichter angepasst werden können.

Im Folgenden werden zwei Ähnlichkeitsmaße kurz umschrieben, um einen Eindruck davon vermitteln zu können, welche Schwierigkeiten auftreten. Das erste Beispiel knüpft an Abschnitt 2.2.3 an, indem ein Wortalignment hergestellt wird, um die Änderungen adäquat zu beschreiben. Als zweites Beispiel wird ein einfaches symmetrisches Abstandsmaß beschrieben.

Wortalignment

Eine Bewertung sollte berücksichtigen, welche Arten von Änderungen vorliegen. Hierzu müssen sie zuerst bestimmt werden. Das ist keine triviale Aufgabe, da Ersetzungen und Umstellungen auch durch eine Kombination von Löschungen und Einfügungen beschrieben werden können und da Zuordnungen nicht eindeutig sind, wenn Wörter doppelt vorkommen.

(2.4) Durch den neuen Bericht wurde der alte Bericht ersetzt.
Der alte Bericht wurde vollständig durch den neuen Bericht ersetzt.

Das konstruierte Beispiel (2.4) ist sicherlich ein Extremfall, der selten vorkommt. Es macht aber deutlich, dass eine Änderung auf unterschiedliche Weise beschrieben werden kann und dass weitere Kriterien notwendig sind, um eine Wahl treffen zu können, welche Beschreibung der Änderungen am angemessensten ist. Mögliche Beschreibungen für die Änderungen im Beispiel (2.4) sind u. a.

a) zwei Umstellung von vier bzw. drei Wörtern und eine Einfügung,

b) zwei Umstellung von drei bzw. zwei Wörtern und eine Einfügung,

c) fünf Umstellungen von Einzelwörtern und eine Einfügung und

d) vier Ersetzungen, eine Löschung und zwei Einfügungen.

Die Beschreibung a) könnte vorgezogen werden, wenn die Kriterien das Zertrennen von Phrasen verbieten. Denkbar wäre zum Beispiel eine Regel, dass eine Wortgruppe nicht zwischen einem Adjektiv und einem Nomen enden kann. Ohne dieses linguistische Wissen wird man vermutlich b) bevorzugen, da hier mehr Wörter unverändert bleiben. „Bericht wurde“ und „Bericht ersetzt“ werden dann als unveränderte Wortgruppen interpretiert. Das in Betracht Ziehen von Wortgruppen und Umstellungen kostet viel Zeit. Wird darauf verzichtet, könnten c) und d) in Frage kommen.

Die Beschreibung der Änderungen hat große Ähnlichkeit mit einem Alignment, das nur zusammenhängende Gruppen erlaubt. Ersetzungen sind Zuordnungen, bei denen die einander zugeordneten Wortgruppen nicht identisch sind. Ein Alignment kann weitere Arten von Änderungen beschreiben als die, die oben erwähnt wurden. Tritt z. B. ein Kompositum in einem Satz getrennt und im anderen zusammen geschrieben auf, dann ist eine 2 : 1 Ersetzung als Beschreibung sinnvoll.

Ein Aligner kann somit die Aufgabe übernehmen, aus den vielen möglichen Beschreibungen der Änderungen eine Beschreibung auszuwählen, die hinsichtlich festzulegender Kriterien optimal ist. Z. B. müssen Ersetzungen deutlich schlechter bewertet werden als Zuordnungen von identischen Wortgruppen, damit Umstellungen, Löschungen und Einfügungen erkannt werden können. Sonst könnte ein Aligner für das Beispiel (2.4) neun Ersetzungen und eine Einfügung als Beschreibung bevorzugen.

Der Aligner könnte auch linguistisches Wissen einsetzen, um die Ersetzungen zu bewerten. Beispielsweise könnten übereinstimmende grammatische Merkmale, die syntaktischen Kategorien der Wortgruppen und der semantische Abstand bewertet werden. Flache Analysen reichen dafür aus: Eine Flexionsanalyse gibt Hinweise darauf, welche grammatischen Merkmale vorliegen. Wortartenfolgen, die ein Tagger (siehe Abschnitt 3.2.3) bestimmen kann, können benutzt werden, um Wortgruppen zu klassifizieren. Für das Nachschlagen der Wörter in einen Thesaurus müssen diese nur auf ihre Grundform reduziert werden.

Die Bewertung eines Wortalignments muss nicht auf die Bewertung der einzelnen Zuordnungen beschränkt bleiben. Weiter oben wurde schon am Beispiel einer Adjektiv-Nomen-Sequenz deutlich, dass die gebildeten Wortgruppen auf ihre linguistische Plausibilität hin überprüft werden sollten. Auch hier können flache Analysestrukuren verwendet werden. Ein so genannter Chunker markiert die Phrasen eines Satzes ohne sie hierarchisch zu ordnen. Die Grenzen der Chunks können mit denen der Wortgruppen verglichen werden.

Die vom Aligner berechnete Bewertung des optimalen Alignments kann nicht ohne Weiteres als Ähnlichkeitsmaß der Sätze verwendet werden. Die Bewertungsfunktion ordnet lediglich die verschiedenen Alignments der zwei vorgelegten Sätze. Die Werte müssen nicht vergleichbar mit den Werten sein, die sich für andere Satzpaare ergeben. Bei dem Entwurf der Bewertungsfunktion muss daher besonders berücksichtigt werden, dass die Bewertungen vergleichbar sein sollen. Alternativ kann das Translation Memory das optimale Alignment mit einer zweiten Bewertungsfunktion beurteilen, die die Änderungen im Hinblick darauf beurteilt, welcher Arbeitsaufwand bei der Anpassung der Übersetzung zu erwarten ist.

Wortpositionen

Eine einfachere Möglichkeit, die Änderungen zu bewerten, bietet die Korrelation der Positionen der Wörter in den zu vergleichenden Sätzen. Tabelle 2.4 zeigt das Prinzip für das

Wort	i	j	$\|i-j\|$	$e^{-\|i-j\|}$
alte	7	2	5	0,007
bericht:1	4	3	1	0,368
bericht:2	8	9	1	0,368
den	2	7	5	0,007
der	6	1	5	0,007
durch	1	6	5	0,007
ersetzt	9	10	1	0,368
neuen	3	8	5	0,007
vollständig	-	5	-	0,000
wurde	5	4	1	0,368
Mittelwert			-	0,151

Tabelle 2.4: Positionsabstände und eine einfache Bewertung

Beispiel (2.4). Die Exponentialfunktion wende ich auf die negativen Differenzen an, damit Wörter, die nur in einem der Sätze auftreten, einfach in die Bewertung integriert werden können. Für sie wird ein unendlicher Positionsabstand angenommen, der zu der Bewertung 0 führt (Zeile „vollständig“ im Beispiel). Je kleiner der Positionsabstand ist, desto größer ist die Bewertung. Die bestmögliche Bewertung 1,0 stellt sich ein, wenn die Wortpositionen identisch sind. Als Gesamtbewertung wird im Beispiel der Mittelwert verwendet.

Weitere Ähnlichkeitsmaße

Baldwin und Tanaka (2000) vergleichen einige Ähnlichkeitsmaße und bieten daher eine gute Übersicht. Sie betrachten sprachunabhängige Maße, die wahlweise die Zeichen oder die Wörter der zu vergleichenden Sätze als Einheiten behandeln. Zu dieser Klasse gehört auch das Maß aus dem vorangehenden Unterabschnitt, da es auch möglich ist, die Zeichenpositionen der einzelnen Buchstaben zu vergleichen. Baldwin und Tanaka beschreiben unter anderem ein auf dem Vector Space Model basierendes Ähnlichkeitsmaß, das im Bereich des Information Retrievals sehr verbreitet ist, das Maß „Editierdistanz“ und zwei Maße, die die Längen der gemeinsamen Zeichen- oder Tokenketten berücksichtigen.

2.3.3 Einsatz flacher Analysestrukturen

Ein Wortalignment bietet sehr viel Spielraum für den Einsatz computerlinguistischer Methoden. Die Wortgruppen können auf linguistische Plausibilität hin geprüft werden, semantische Netze können eingesetzt werden, um die Ähnlichkeit unterschiedlicher Wörter zu messen und syntaktische Strukturen können ein hierarchisches Alignment induzieren. Aber auch ohne Wortalignment läßt sich linguistisches Wissen in ein Ähnlichkeitsmaß integrieren. Dazu folgen einige Beispiele.

Carl und Hansen (1999) berücksichtigen bei der Bewertung nur die Grundformen der Wörter. Voraussetzung dafür ist, dass das Referenzmaterial und der Anfragesatz auf mit Grundformen annotiert sind. Das resultierende System wird von Carl und Hansen lexembasiertes TM, kurz LTM genannt.

Viele Ähnlichkeitsmaße, die für das Erstellen von Satzalignments zwischen verschiedensprachigen Texten entworfen wurden, lassen sich für den monolingualen Einsatz anpassen. Beispielsweise nutzt das Maß von Piperidis et al. (2000) nur Wortarteninformationen aus. Sie bilden eine Linearkombination der Häufigkeiten einiger Wortarten im Ausgangssatz und vergleichen diese Zahl mit der Anzahl der Wörter aus offenen Wortklassen im Zielsatz. (Seite 121–124) Dies kann ohne Änderung für Sätze einer Sprache vorgenommen werden. Problematisch ist nur die Wahl der Gewichte der Linearkombination. Wenn keine Sätze als Trainingsmaterial vorliegen, die trotz unterschiedlicher Wortarthäufigkeiten den gleichen Inhalt haben, gibt es keinen Grund, Gewichte ungleich eins zu wählen.

Planas und Furuse (2000) unterteilen das Referenzmaterial in mehrere Ebenen, die sie TELA-Ebenen nennen. Die einfachste Ebene enthält den Text als Zeichenfolge. Dann folgt eine Ebene, in der die Wörter isoliert sind. Weitere Ebenen speichern Schriftauszeichnungen, Informationen für die Indexverwaltung und Verweise. Darüber hinaus gibt es abgeleitete Ebenen, die flache Analysestrukturen enthalten. Abgeleitet bedeutet, dass sie jederzeit neu bestimmt werden können, nämlich durch die zugrunde liegende Analyse. Diese Ebenen annotieren Grundformen, Wortarten und unstrukturierte Phrasen, so genannte Chunks. Planas und Furuse skizzieren ein Matching-Verfahren, das je Wortposition die spezifischte Ebene ermittelt, auf der eine Übereinstimmung gefunden werden kann. Ein Beispiel ist angegeben, in dem „NTT really stayed strong Monday.“ und „Sony stayed stronger Tuesday.“ verglichen werden. Das erste und letzte Wort stimmen nur in der Wortart überein. Das zweite Wort wurde gelöscht. Dann folgt ein übereinstimmendes Wort. An der vorletzten Position stimmen die Grundformen, aber nicht die Wörter überein. Diese Informationen können benutzt werden, um die Unterschiede im Fuzzy-Match zu markieren. Planas und Furuse (2000) haben jedoch eine Anwendung in der automatischen Übersetzung im Blick und erlauben zur Vereinfachung der Berechnung der Matches keine Einfügungen und Ersetzungen. Letzteres ist keine starke Einschränkung, da eine Ersetzung erst vorliegen würde, wenn an der Wortposition alle TELA-Ebenen nicht übereinstimmen.

2.3.4 Verwendung der Übersetzungsvorschläge

Translation Memorys bieten gewöhnlich zwei Arten an, wie sie dem Benutzer Übersetzungsvorschläge unterbreiten. Zum einen kann der am besten bewertete Vorschlag ohne Nachfrage in den Editierbereich, in dem die Übersetzung verfasst wird, als Vorlage eingefügt werden. Eventuell wird annotiert, dass es sich um einen Fuzzy-Match handelt, damit der Übersetzer den Vorschlag gründlicher prüft als einen Übersetzungsvorschlag, der auf einem Exact-Match zurück geht.

Zum anderen kann der Übersetzer eine Liste aller Fundstellen abrufen, die nach der berechneten Relevanz geordnet ist. Auf der ausgangsprachlichen Seite können für jede Fundstelle die Unterschiede zu dem zu übersetzenden Satz hervorgehoben werden. Ein in der Bewertungsphase erstelltes Wortalignment ist hierfür eine idelae Grundlage. Die jeweiligen Übersetzungen werden ohne jede Hervorhebung mit angegeben und können vom Übersetzer als Vorlage für die zu erstellende Übersetzung ausgewählt werden. Mir ist nicht bekannt, ob die Wahl von Übersetzern anhand der Ausgangstexte oder der Übersetzung getroffen wird. In letzteren Fall könnte es nützlich sein, die Übersetzungsvorschläge so darzustellen, dass einander ähnliche Vorschläge leicht erkannt werden können.

Übersetzungsvorschläge, die auf Fuzzy-Matches basieren, erfordern i. d. R. Anpassungen. (Ausnahmen ergeben sich z. B. beim Übersetzen einer korrigierten Fassung eines Textes, der in einer Rohfassung, die viele Fehler enthält, bereits übersetzt wurde.) Ein Translation Memory bietet grundsätzlich nur Übersetzungsvorschläge in der Form an, wie es sie im Referenzmaterial vorfindet. Selbst einfache Anpassungen, wie z. B. das Ersetzen

von Produktbezeichnungen, Datumsangaben oder Zahlen, muss der Übersetzer vornehmen.

2.4 Layout-Information

Ein Dokument ist mehr als eine Abfolge von Wörtern. In Abschnitt 2.2.1 wurde bereits erwähnt, dass ein Absatzende besonders vermerkt wird. Auf ähnliche Weise sind Überschriften, Listen und viele andere Elemente ausgezeichnet. Sowohl strukturelle Informationen, z. B. ob es sich um eine Kapitel- oder Abschnittsüberschrift handelt, als auch konkrete Anweisungen zur Darstellungen, wie die zu verwendende Schrift und Abstände, können annotiert sein.

2.5 Evaluationkriterien

Es ist schwierig, Kriterien für den Vergleich von Translation Memory Systemen zu finden. Maßstab soll sicherlich sein, wie gut das Translation Memory dem Übersetzer hilft, seine Arbeit auszuführen. Die Güte der Hilfe kann an der Zeitersparnis[11] gemessen werden, wenn davon ausgegangen werden kann, dass die Übersetzungsqualität unverändert bleibt. Ansonsten muss die Qualität mit in die Bewertung einbezogen werden. Wenn die Übersetzungsqualität mit brücksichtigt wird, dann können Translation Memorys nicht nur untereinander, sondern auch mit anderen CAT Systemen verglichen werden. Des Weiteren wird häufig behauptet (vergleiche (Seewald-Heeg und Nübel, 1999, Seite 119)), dass Translation Memorys die Übersetzungsqualität steigern, da sie die Konsistenz der Übersetzungen erhöhen. Andererseits könnte die Qualität auch durch Fehlübersetzungen leiden. (Webb, 1998, Abschnitt 9) weist darauf hin, dass die Benutzung eines Translation Memorys die Zahl der Nachbearbeitungszyklen reduzieren kann.

Der Aufwand, vergleichbare Texte unter gleichen Bedingungen zu übersetzen, ist sehr hoch. Trotz des Aufwands dürfte es schwierig sein, die Ergebnisse zu reproduzieren, da die gemessenen Werte von den Übersetzern abhängen, die für den Test eingesetzt werden. Man wird also möglichst einen anderen Weg suchen, um ein Translation Memory zu evaluieren.

Eine erste Vereinfachung wäre, dass man Sätze, für die das Translation Memory keinen Übersetzungsvorschlag unterbreitet, nicht vom Übersetzer bearbeiten läßt, sondern eine pauschale Dauer für die Übersetzungstätigkeit ansetzt, z. B. von 15 Sekunden je Wort. Ebenso muss der Übersetzer nicht bemüht werden, wenn ein Übersetzungsvorschlag korrigiert werden muss, der bereits von einem anderen Translation Memory zur gleichen Textstelle unterbreitet wurde. Hier kann die Dauer der erstmaligen Korrektur unterstellt werden.[12] Im Falle von Exact-Matches wird häufig vereinfachend davon ausgegangen, dass die Übersetzungsvorschläge immer richtig und keine Nachbearbeitungen notwendig seien. Mögliche Ambiguitäten oder Kontextabhängigkeiten werden ignoriert. Eine wesentlich weitergehende Vereinfachung wäre, ganz auf die Messung der Übersetzungsdauer zu verzichten und diese nur abzuschätzen. Somers (1999) berichtet im Zusammenhang mit der Evaluation von MT Systemen, dass es üblich ist, die Übersetzungsvorschläge mit einer Musterübersetzung zu vergleichen (Seite 145–146). Diese Art der Evaluation bietet den Vorteil, dass kein Übersetzer benötigt wird, wenn Testtext und Musterübersetzung dem Referenzmaterial entnommen werden. Zwar geht es bei Somers (1999) um die Messung

[11] Für die Nützlichkeit im Berufsalltag spielen natürlich auch andere Faktoren eine Rolle, insbesondere die Akzeptanz des Systems. Diese können aber nur mit wesentlich höheren Aufwand evaluiert werden.

[12] In beiden Fällen wird vereinfachend davon ausgegangen, dass die Sätze isoliert, also unabhängig vom Kontext übersetzt werden können.

der Qualität der Übersetzung und nicht um die Dauer der Erstellung. Aber als verwendete Vergleichsmethoden werden selbst solche genannt, die die Anzahl der notwendigen Editierschritte zählen, um den Übersetzungsvorschlag in die Musterübersetzung zu überführen. Diese Zahl ist ein gutes Maß für die Dauer der Änderung. Carl und Hansen (1999) benutzen ein solches maschinelles Translation Score, um verschiedene Systeme zu vergleichen.

Durch die direkte Bewertung der Übersetzungsvorschläge wird die Benutzerschnittstelle aus der Evaluation ausgeblendet. Das heißt, dass die Art, wie die Übersetzungsvorschläge dem Benutzer des Translation Memorys präsentiert werden, keine Rolle spielt. Beim Vergleich verschiedener Evaluationen muss man daher nicht nur berücksichtigen, welches Referenzmaterial und welcher Ausgangstext verwendet wurden, sondern ebenso prüfen, was genau evaluiert wurde.

Häufig gibt es gute Gründe, warum einzelne Komponenten ausgeblendet werden. Wer nur eine einzelne Komponente entwickelt, möchte sie mit den entsprechenden Komponenten anderer Systeme vergleichen. Für Translation Memorys können folgende Komponenten identifiziert werden:

- Alignment des Referenzmaterials,
- Auswahl und Bewertung der Übersetzungsvorschläge,
- Präsentation der Übersetzungsvorschläge und
- Integration in den Editor.

Aus computerlinguistischer Sicht sind besonders die ersten beiden Komponenten interessant, da hier Methoden des Fachs im Vordergrund stehen. Dennoch bieten auch die anderen Komponenten Raum für computerlinguistische Anwendungen.

Somers (1999) nennt zur Evaluation der Übersetzungsvorschläge verschiendene Ähnlichkeitsmaße (vergleiche Abschnitt 2.3) und die Bewertung durch Spezialisten (Seite 147–148). Beispielsweise bitten Cranias et al. (1994) fünf Übersetzer, alle Vorschläge in vier vorgegebene Nützlichkeitsklassen einzuteilen. In der Auswertung werden die Anzahlen je Klasse einfach summiert. Der Grad der Übereinstimmung der einzelnen Bewertungen wird nicht berechnet. Hierfür wäre die Kappa-Statistik geeignet. Eine leicht verständliche Einführung findet sich in Carletta (1996). Auch werden in einigen Arbeiten die Ergebnisse verschiedener, maschineller Bewertungen angegeben ohne die Unterschiede genauer zu untersuchen.

Ganz andere Evaluationkriterien, die die Bedürfnisse des beruflichen Übersetzers im Blick haben, werden in der Hausarbeit von Erpenbeck et al. (2000) genannt. Die Autoren stützen sich dabei wesentlich auf die Empfehlungen der EAGLES-Kommission, die auch in Seewald-Heeg und Nübel (1999) und Reinke (1999) verwendet werden. Ebenfalls klar als Produktevaluation angelegt sind die Kriterien, die im ARG-Projekt[13] „Computer-Assisted Translation for Irish“ zur Evaluation von vier Produkten benutzt werden. Die Arbeit von Feder (2001) konnte hier leider nicht mehr berücksichtigt werden.[14]

2.5.1 Produkte

Während der Recherchen für diese Arbeit konnten viele Hinweise auf Produkte und Hersteller gefunden werden. In Tabelle 2.5 sind diese Informationen zusammengestellt. Man beachte, dass Produktbezeichnungen und Hersteller sich geändert haben oder vom Markt

[13] `http://www.compapp.dcu.ie/~kkeogh/`

[14] Titel und Bibliographie sind vielversprechend.

Hersteller	Produkt
Alchemy	Catalyst
Alpnet	Joust / TSS (Translation Support System)
Atril	DejaVu
ESTeam	ESTeam Translation Memory
Eurolang	Optimizer
IBM	TranslationManager
linguatec	Personal Translator 2000
MorphoLogic	MoBiMem
SDL	SDLX
STAR	Transit
Trados	Translator's Workbench
Zeres	Zeresztrans

Tabelle 2.5: einige Translation Memory Produkte

verschwunden sein können. Der Leser möge diese Liste als Ausgangspunkt für eigene Recherchen nutzen. In dieser Arbeit wird auf die Marktsituation nicht weiter eingegangen. Arbeiten, die sich mit Produkten beschäftigen, sind Dennett (1995), die Seminararbeit von Erpenbeck et al. (2000) und der von Language Automation Inc. verbreitet Text, der im Literaturverzeichnis unter Unbekannt (2001) gelistet ist.

2.6 Zusammenfassung

Ein Translation Memory ermöglicht die Wiederverwendung bereits erstellter Übersetzungen. Für Sätze, zu denen ein indentischer oder ähnlicher Satz im Referenzmaterial gefunden werden kann, präsentiert es Übersetzungsvorschläge, die im Referenzmaterial belegt sind und somit in sich korrekt sind, wenn das Material auf der zielsprachlichen Seite keine Fehler enthält.

Prinzipielle Schwächen eines Translation Memorys sind, dass nicht immer ein ausreichend guter Fuzzy-Match zur Verfügung steht, dem ein Übersetzungsvorschlag entnommen werden könnte, und dass gefundene Übersetzungen im neuen Kontext unpassend sein können. Darüber hinaus muss das Referenzmaterial dem für die Übersetzung gewünschten Stil und Genre entsprechen, damit ein Translation Memory adequate Vorschläge unterbreiten kann.

In diesem Kapitel lag der Schwerpunkt auf die Darstellung der Funktionsweise eines Translation Memorys. Zwei Phasen sind zu unterscheiden: Zur Vorbereitung der Arbeit mit einem Translation Memory wird ein Satzalignment für das Referenzmaterial erstellt. Dies ist Voraussetzung dafür, während der Übersetzungstätigkeit schnell und zuverlässig auf die Übersetzung von relevantem ausgangssprachlichen Material zugreifen zu können. Die zweite Phase ist die Anwendungsphase. Das Translation Memory unterstützt den Übersetzer, indem es Übersetzungsvorschläge unterbreitet, die dem Referenzmaterial entnommen wurden. Die relevanten Stellen zum zu übersetzenden Satz werden mit einem Ähnlichkeitsmaß identifiziert. Ein Ähnlichkeitsmaß ordnet die Kandidaten für die Fuzzy-Matches und wird verwendet um zu entscheiden, welche Sätze als Fundstelle akzeptiert

werden.

Beide Bereiche, Satzalignment und Ähnlichkeitsmaß, bieten viel Freiraum für den Einsatz computerlinguistischer Methoden. Zum Satzalignment wurde hier nur das Grundprinzip erläutert, da es nicht direkt in die automatischen Erstellung von Übersetzungsvorschlägen eingebunden ist, sondern zur Aufbereitung des Referenzmaterials als linguistische Ressource dient.[15] Das Ähnlichkeitsmaß bestimmt dagegen die Übersetzungsvorschläge des Translation Memorys. Wenn auf Fuzzy-Matches zurückgegriffen werden muss, entscheidet das Ähnlichkeitsmaß, welche Stellen im Referenzmaterial dem Übersetzer präsentiert werden. Es wurden daher mehrere Möglichkeiten skizziert, wie ein Wert für die Ähnlichkeit bestimmt werden kann. Am umfangreichsten dargestellt wurde das Wortalignment, da in dessen Bewertung verschiedene linguistische Analysen einfließen können.

[15] Den Nutzen linguistischen Wissens beim Erstellen von Satzalignments hat Tschorn (2002) in seiner Magisterarbeit untersucht.

Kapitel 3

Korpusaufbereitung für CAT-Systeme

In dieser Arbeit wird ein bilinguales Korpus verwendet, um Fallbeispiele für die Betrachtung einzelner Probleme der datengestützten Übersetzung untersuchen zu können. Das verwendete Korpus besteht aus einer Sammlung von deutschen und englischen Texten zusammen mit ihren jeweiligen englischen und deutschen Übersetzungen. Ein Teilkorpus kann auch Paare von deutschen und englischen Texten enthalten, die aus einer dritten Sprache übersetzt wurden.

In diesem Kapitel wird das Korpus vorgestellt. Zuerst werden die Quellen genannt. Dann werden die Schritte der Vorverarbeitung beschrieben, die das Korpus in eine Form bringen, in der es leichter genutzt werden kann. Anschließend werden kurz einige quantitativen Abgaben zum Korpus gemacht. Schließlich wird eine Stichprobe aus dem Korpus vorgestellt, die zeigen soll, welche Arten von Fuzzy-Matches erwartet werden können und wie häufig sie auftreten. Die Stichprobe wird Grundlage für die Betrachtungen im Kapitel 4 sein.

3.1 Studienprojekt KoKS

Im Studienprojekt KoKS wurde ein bilinguales Korpus aufgebaut, das wie in einem Translation Memory auf Satzebene alignt ist. Es kann daher ohne große Anpassungen in dieser Arbeit verwendet werden. Die Nutzung des Korpus wird wesentlich dadurch erleichtert, dass der Autor selbst Projektmitglied war und mit den Datenformaten und Werkzeugen vertraut ist, die im KoKS-Projekt entwickelt wurden.

Am Studienprojekt KoKS nahmen insgesamt sechs Studenten des Studiengangs Computerlinguistik und Künstliche Intelligenz teil. Die geplante Dauer betrug ein Jahr. Sie konnte aber nicht eingehalten werden. Das Projekt erstreckte sich von Oktober 2000 bis Januar 2002.

Die Projektergebnisse sind in einem 641 Seiten umfassenden Abschlussbericht von Erpenbeck et al. (2002) dokumentiert, der in einer um die Sitzungsprotokolle und persönliche Schilderung der Projekterfahrungen gekürzten Fassung öffentlich auf der Projektwebseite zugänglich ist.

3.1.1 Kollokationen

Der Name KoKS steht für Korpusbasierte Kollokationssuche. Im KoKS-Projekt sollte ein System entwickelt werden, dass Kollokationen aus einem bilingualen Korpus extrahiert. Kollokationen sind Mehrwortausdrücke oder Phrasen, in denen nicht jedes Wort durch ein Synonym ersetzt werden kann.[1] Beispielsweise kann in „ins Gras beißen" das Nomen nicht ersetzt werden. „In den Wiesenbewuchs beißen" hat nicht die Bedeutung „sterben". (Zu dieser Bedeutung kann man zwar gelangen, indem man „um die Ecke denkt". Aber bei der „Ecke" handelt es sich um die Kollokation „ins Gras beißen".)

Die im KoKS-Projekt verwendete Definition von Kollokationen ist spezifischer als die hier dargestellte, ist aber für diese Arbeit jedoch nicht wichtig. Kollokationen sind beim Übersetzen nur insofern interessant, als dass sie besondere Aufmerksamkeit erfordern. Sie können nicht kompositionell, d. h. nicht jeder Bestandteil kann unabhängig vom Kontext, übersetzt werden. Natürlich kann man einwenden, dass es für eine gute Übersetzung der Regelfall ist, dass der gesamte Kontext Einfluss auf die Wortwahl hat.

Im KoKS-Projekt wurde versucht, Kollokationen daran zu erkennen, dass ihre Übersetzung nicht mit vorhandenen Wörterbucheinträgen erklärt werden kann. Dazu verwendet das KoKS-System ein Abstandsmaß, das den Grad der Übereinstimmung von Ausgangsphrase und Übersetzung mit Hilfe eines Wörterbuchs misst. Ergebnisse haben Kummer und Wagner (2002) vorgestellt.

3.1.2 Korpusquellen

Von dem im KoKS Projekt zusammengestellten Korpus wurden nur die zwei Teilkorpora „DE-News" und „EU" übernommen. Die übrigen Teilkorpora wurden entweder bereits im KoKS Projekt aus verschiedenen Gründen (siehe Abschlussbericht) nicht weiter verwendet oder ihre Berücksichtigung erschien wegen ihres geringen Umfangs nicht lohnenswert. Im KoKS-Abschlussbericht werden als Quelle der übernommenen Teilkorpora „De-News" und „EU" die Webseiten `http://www.isi.edu/~koehn/publications/de-news/` und `http://europa.eu.int/rapid/start/welcome.htm` genannt.

Kummer und Wagner (2002) haben für ihre Untersuchung zusätzlich die ersten vier Harry Potter Bücher von Joanne K. Rowling als literarischen Teilkorpus erschlossen, da sie hofften, dort eine höhere Dichte von Kollokationen vorzufinden. Die Werke wurden im Juni 2002 über das Internet aus nicht notierten Quellen bezogen. Es ist davon auszugehen, dass Unbekannte die Bücher eingescannt und mit einer OCR Software in Text oder PDF umgewandelt haben. Die englischen und deutschen Fassungen konnten innerhalb zweier Tage zusammengestellt werden. Probleme bereiteten ein Teil der PDF-Dokumente. Wenn die Extraktion des Textes nicht gelang, musste eine weitere Quelle gefunden werden.

3.2 Vorverarbeitung

Die Dokumente des Korpus müssen einige Vorverarbeitungsschritte durchlaufen, bevor sie in den Programmen des KoKS-Projekts und den für diese Arbeit speziell erstellten Softwarewerkzeugen verwendet werden können. Die Vorverarbeitung ist bis auf die zusätzliche Indizierung mit der des KoKS-Projekts identisch. Die einzelnen Schritte beschreiben Erpenbeck et al. (2002) im KoKS-Abschlussbericht ausführlich. Hier ist die Darstellung

[1] Es gibt andere Verwendungsweisen des Begriffs. Sehr verbreitet ist auch eine rein statistische Sichtweise, nach der jede Wortverbindung eine Kollokation ist, die häufiger auftritt, als dies von den einzelnen Häufigkeiten der beteiligten Wörter zu erwarten wäre.

<H1>
Mein Wochenende
</H1>
Letztes Wochenende war langweilig. Die Fete zum Ferienbeginn fiel ins Wasser, weil die Disco abgebrannt war. Ausserdem kam auch nichts Anstaendiges im Fernsehn.

<H1>
My weekend
</H1>
Last weekend was boring. The school's out party was called off. The club had burned down. Also, there was nothing on the telly.

Abbildung 3.1: Aufbereitetes Dokumentpaar

knapper gehalten und richtet sich vor allem auf Aspekte, die für diese Arbeit relevant sind oder im KoKS-Abschlussbericht nicht behandelt werden.

Ziel der Vorverarbeitung ist eine einheitliche Speicherung der Dokumente und zusätzlicher Information, die für die Anwendung relevant sind, wie z. B. das Satzalignment, das sowohl im KoKS-System als auch in dieser Arbeit Ausgangspunkt für jede Weiterverarbeitung ist. Während beim KoKS-System zusätzlich die Annotation der Wortarten im Vordergrund stehen, spielen in dieser Arbeit flexible Suchmöglichkeiten eine wichtigere Rolle.

3.2.1 Aufbereitung und Normalisierung

Die Dokumente, aus denen sich das KoKS-Korpus zusammen setzt, stammen aus verschiedenen Quellen. Entsprechend vielfältig sind die Probleme, die bei der Zuordnung der deutschen und englischen Fassung eines Dokuments auftraten. Mit computerlinguistischen Methoden konnten diese im KoKS-Projekt gelöst werden: Zur Sprachidentifikation wurden Häufigkeitsverteilungen der auftretenden Buchstaben-n-Gramme gemessen, und zur Überprüfung des Dokumentalignments einer Quelle wurde der KoKS-Aligner in einer modifizierten Fassung eingesetzt.

Nach dieser Aufbereitung liegen die Dokumente in einem Verzeichnisbaum und werden durch eine XML-Datei (`index.xml`) je Teilkorpus beschrieben. Die Beschreibung schließt die Zuordnung der deutschen und englischen Fassungen ein. I. d. R. wurden zusätzlich die Dateinamen für die Dokumente so gewählt, dass Dokumente, die Übersetzungen voneinander sind, durch ein Präfix erkannt werden können. Abbildung 3.1, aus der Abschlusspräsentation des KoKS-Projekts adaptiert wurde, zeigt ein sehr kurzes Dokumentpaar, das im folgenden verwendet wird, um die einzelnen Vorverarbeitungsschritte zu illustrieren.

Im zweiten Vorverarbeitungsschritt werden die Formate der Dokumente normalisiert, um in den weiteren Schritten ein einheitliches Format voraussetzen zu können. Für jedes Dateiformat, das in einer Korpusquelle verwendet wird, steht ein Normalisierungsmodul bereit, das Dokumente auf eine Abfolge von Überschriften und Absätzen reduziert und sämtliche Layout- und sonstige Strukturinformationen entfernt. Dies ist ein Unterschied zu gewöhnlichen Translation Memorys. Dort bleiben die Formatanweisungen erhalten, sodass ein Exact-Match nur möglich ist, wenn auch die Formatierungen übereinstimmen. Im Translation Memory dieser Arbeit werden Formatierung beim Matching nicht berücksichtigt, da sie nicht gespeichert sind.

Die Normalisierung fügt nach Absätzen und Überschriften eine Markierung ein. Mar-

Mein Wochenende	My weekend
<ABSATZ>	<ABSATZ>
Letztes Wochenende war langweilig. Die Fete zum Ferienbeginn fiel ins Wasser, weil die Disco abgebrannt war. Ausserdem kam auch nichts Anstaendiges im Fernsehn.	Last weekend was boring. The school's out party was called off. The club had burned down . Also, there was nothing on the telly.
<ABSATZ>	<ABSATZ>

Abbildung 3.2: Normalisiertes Dokumentpaar

kierungen werden in spitzen Klammern gesetzt, da sie dann im nachfolgenden Vorverarbeitungsschritt keine Probleme bereitet, siehe Abbildung 3.2. Die Ähnlichkeit zu SGML-Markierungen verleitet dazu, anzunehmen, es handle sich um eine Startmarkierung. Die Markierung zeigt hier aber das Ende eines Absatzes (oder einer Überschrift) an.

In den weiteren Schritten wird nicht zwischen Überschriften und Absätzen unterschieden. Überschriften sind im KoKS-System spezielle Absätze, die gewöhnlich ohne Satzzeichen oder mit Frage- oder Ausrufungszeichen enden und nicht mehr als einen Satz enthalten. Nach der Normalisierung spielt Whitespace[2] außer als Worttrenner keine Rolle mehr. Abbildung 3.2 zeigt das normalisierte Beispiel. Man beachte, dass der Punkt nach „burned down" abgerückt ist. Der SGML-Parser wird offenbar nicht korrekt benutzt. Beim HTML-Normalisierungsmodul tritt dieser Effekt nicht auf. Dies ist aber kein akutes Problem, da in dem KoKS-Korpus Formatierungen selten oder gar nicht auftreten.

Aufbereitung des Harry-Potter Korpus

Beim Harry-Potter Korpus, das erst nach dem Ende des KoKS-Projekts von Norman Kummer und dem Autor dieser Arbeit erschlossen wurde, mussten die Dokumente in kleinere Dateien zerlegt werden, da sich der KoKS-Aligner in Laufzeit und Speicherplatzbedarf nicht besser als quadratisch zur Satzanzahl verhält. Die vollständigen Bücher, die jeweils zwischen ca. 6 500 und 15 000 Sätze umfassen, sind für den Aligner zu groß. (Zur Arbeitsweise des Aligners siehe Abschnitt 3.2.5 weiter unten.)

Die Aufteilung muss in der deutschen und englischen Fassung an sich entsprechenden Stellen erfolgen, damit die resultierenden Dokumente Übersetzungen voneinander bleiben. Hierzu wurden die beiden Sprachfassungen in zwei Texteditoren geöffnet und an geeigneten Stellen Trennzeilen eingefügt, an denen die Texte anschließend in Einzeldateien aufgeteilt wurden.

Ein weiteres spezielles Problem des Harry-Potter Korpus ergibt sich daraus, dass die Dokumente per OCR von einer Buchvorlage erfasst wurden. Der Text wird daher in regelmäßigen Abständen durch Seitenzahlen unterbrochen, und Zeichen können falsch erkannt sein. Die Zeilen, die die Seitenzahlen enthalten, wurden mit einem Suchmuster identifiziert und entfernt. Weil die Erkennungsqualität der OCR bei den Seitenzahlen sehr schlecht war, mussten neben Ziffern auch weitere Zeichen, wie „l" und „*" in das Suchmuster aufgenommen werden. Möglicherweise wurden dadurch einige zum Text gehörende

[2] Sammelbezeichnung für Elemente einer Zeichenfolge (String), die den Fluss der Zeichen unterbrechen, z. B. Leerzeichen, Zeilenumbruch, -vorschub, Seitenwechsel und Tabulatoren.

Text	erwartete Tokenanzahl	KoKS-Tokenanzahl
John O'Brien	2	2
award-winning	1	1
film and television.	4	4
John's other television credits include	5	6
'Water Rats'	5	5
I've done five	4	4
That's nice.	4	4
Abk. f. Abkürzung	3	3
von Sätzen usw. Der Name steht	?	6

Tabelle 3.1: Schwierigkeiten bei der Tokenisierung

Zeilen gelöscht.[3]

Zu erwarten wäre, dass die Silbentrennung der gedruckten Vorlage einen so großen Teil der Wörter zertrennt, dass die meisten Sätze betroffen sind. Jedoch sind innerhalb der einzelnen Seiten Wörter am Zeilenende nur extrem selten getrennt. Da dagegen am Seitenende Wörter häufig getrennt sind, ist dies vermutlich kein Merkmal der Bücher, sondern erklärt sich als nachträgliche Korrektur derjenigen, die die Texte im Internet verbreiten, oder als automatische Anpassung durch die OCR-Software. Während getrennte Wörter für das KoKS-System nur eine höhere Quote unbekannter Wörter zur Folge haben, verringern sie in der Translation Memory Anwendung beim Fuzzy-Matching die Ähnlichkeit zum Anfragesatz unnötig.

3.2.2 Tokenisierung

Vor der Tokenisierung sind die Dokumente Zeichenfolgen, die nur gelegentlich von Absatzendemarkierungen unterbrochen werden. Die Tokenisierung legt fest, welche Zeichenfolgen in der weiteren Verarbeitung als eine Einheit betrachtet werden. Die Einheiten werden Token genannt, was selbst soviel wie Zeichen[4] bedeutet. Damit soll betont werden, dass sie immer nur als ganzes verarbeitet werden. Token sind gewöhnlich Wörter oder Zahlen. Häufig können sie am sie umgebenen Leeraum erkannt werden. Eine gute Tokenisierung einer längeren Zeichenfolge ist aber nur in Ausnahmefällen identisch mit einer einfachen Zerlegung der Eingabe an Leerzeichen. So bilden z. B. Satzzeichen keine Einheit mit dem vorangehenden Wort. Sie werden entweder als eigenes Token behandelt oder ganz ignoriert. Der im KoKS-System verwendete Tokenisierer behält Satzzeichen bei.[5] Weitere Sonderfälle stellen Klammern, Bindestriche und Anführungszeichen dar. Tabelle 3.1 zeigt einige problematische Textfragmente, die größtenteils einem ABC Online Interview entnommen wurden, und die Anzahl der Token. Abkürzungen am Satzende absorbieren beim

[3]Darüber hinaus wurden einige Zeilen, vor allem Überschriften, die wegen vieler OCR-Fehler unleserlich waren, absichtlich entfernt. Um das Satzalignment nicht zu erschweren, wurden auch die entsprechenden Passagen in der anderen Spachfassung herausgenommen. Hiervon ist aber nicht das gesamte Korpus betroffen, da einer der beiden menschlichen Aufbereiter diese Löschungen ablehnte.

[4]Im Unterschied zur Menge der Zeichen ist die Menge der Token nicht endlich.

[5]Wenn in dieser Arbeit von der Anzahl der Token oder Wörter (z. B. eines Satzes) die Rede ist, sind also Satzzeichen mitgezählt.

KoKS-Tokenisierer den Punkt, der dann nicht mehr als eigenes Token zur Verfügung steht.[6]

Die Tokenisierung ist im KoKS-System kein eigenständiges Modul, sondern wird zusammen mit dem POS-Tagging (siehe unten) vom IMS TreeTagger ausgeführt. Zwar können die einzelnen Komponenten des IMS TreeTaggers nicht angepasst werden. Aber zwischen ihnen kann die Ein- und Ausgabe manipuliert werden. Im KoKS-Projekt wurde davon Gebrauch gemacht, um das Verhalten bei Punkten zu ändern. Nicht jeder Punkt ist automatisch ein Satzzeichen. Punkte treten in Abkürzungen, Zahlen und Nummerierungen auf. Der IMS Tagger setzt eine Liste von Abkürzungen ein, um Punkte unterschiedlich zu behandeln. Wird nach einem Punkt klein geschrieben, dann wird der Punkt anscheinend grundsätzlich zum vorangehenden Token gezählt.

Manning und Schütze (1999) diskutieren weitere Probleme der Tokenisierung (Seite 124–131). U. a. ist die Situation bei Klitika im Englischen komplizierter, als in der Tabelle 3.1 dargestellt. Ein Problemfall von mehreren ist das Possessivum im Plural, wie in „the boys' toys".

Anpassung der Schreibung

Die zweite KoKS-Erweiterung des IMS Taggers betrifft die Orthographie. Ein Teil der Dokumente verwendet keine Umlaute und Eszett. Vor den weiteren Vorverarbeitungsschritten müssen diese Wörter korrigiert werden. Dazu werden Regeln und die Vollformenliste der bereits verarbeiteten Dokumente verwendet.

Mit dem Harry-Potter Korpus stellt sich die neue deutsche Rechtschreibung als weiteres Problem heraus. Die beiden häufigsten betroffenen Wörter „dass" und „muss" sollten eigentlich durch die Umlaut- und Eszettkorrektur angepasst werden. Dies geschieht aber nicht, da die Vollformenliste die Wörter auch in der neuen Schreibung enthält. Mit der Absicht eine korrekte Vollformenliste aufzubauen wurden zuerst die Wörterbücher und Teilkorpora verarbeitet, die keine Umlaut- und Eszettkorrektur erfordern. Dann wurde das Korrekturmodul aktiviert und die restliche Teilkorpora verarbeitet. Da das Ziel die Korrektur der Teilkorpora war, die keine Umlaute und Eszett verwenden, wurde nicht beachtet, dass eines der Wörterbücher die neue Rechtschreibung verwendet.[7] Warum nicht bei der Überprüfung der Ausgabe des Korrekturmoduls aufgefallen ist, dass die häufigen Wörter „dass" und „muss" weiterhin auftreten, lässt sich nicht mehr rekonstruieren.[8]

Analog könnte die im vorangehenden Abschnitt erwähnte Silbentrennung an Zeilenumbrüchen von einem Tokenisierer entfernt werden. Eine Überprüfung, ob die verschmolzenen Wörter bereits im System bekannt sind, könnte verhindern, dass Gedanken- oder Bindestriche, die zufällig am Zeilenende stehen, als Trennstrich bewertet werden. Dies wäre ein Beispiel dafür, dass Whitespace nicht immer Token trennt. Der KoKS-Tokenisierer leistet dies jedoch nicht.

[6] In KoKS kann das Satzende trotzdem repräsentiert werden, da eine Tokenfolge von Markierungen (analog zu Absatzendemarkierung) unterbrochen werden kann. Beispiele hierzu finden sich im Abschnitt 3.2.3, siehe Abbildung 3.3.

[7] Betroffen ist das Wörterbuch mit der KoKS-internen Bezeichnung `wb1`. Es scheint vollständig in der neuen Rechtschreibung verfasst zu sein und enthält neben Einzelworteinträgen auch Phrasen wie z. B. „leider muss ich sagen" und „zu der Anschauung gelangen, dass". Das Wörterbuch `wb1` sollte in Zukunft nicht zum Aufbau der initialen Vollformenliste verwendet werden. Das gleiche gilt für das Wörterbuch `wb3`, das die alte Rechtschreibung benutzt, aber viele falsche Umlaute, z. B. „daürnd" und „Baumverhaü", enthält, und zwar bereits in der Rohfassung. Eventuell lohnt es sich, hier sämtlich Umlaute mit „ue" usw. auszuschreiben und dann die KoKS-Umlautkorrektur anzuwenden.

[8] Das Projektmitglied, das diese Prüfung vorgenommen hat, berichtete, dass nach der Korrektur mehr Wörter korrekt seien als zuvor. Dass Wörter mit Umlaut wesentlich häufiger auftreten als „dass" und „muss" zusammen, könnte erklären, warum letztere Wörter keine Aufmerksamkeit fanden. Eine andere Erklärung könnte sein, dass vielleicht eine andere Vollformenliste verwendet wurde.

Unumkehrbarkeit

Im Allgemeinen ist die Tokenisierung nicht umkehrbar. Zur Ausgabe von Text bietet es sich an, die Token leerzeichengetrennt aneinander zu hängen und Leerzeichen vor Satzzeichen und schliessenden Klammern und nach öffnenden Klammern zu löschen. Bei nicht typographischen Anführungszeichen ist die Situation schwieriger. Hier kann nur mit größerem Aufwand entschieden werden, welches Leerzeichen unerwünscht ist. Es kann aber nicht garantiert werden, dass das Resultat mit dem ursprünglichen Text identisch ist, da der Tokenisierer nicht entsprechend entworfen wurde. Dies wird an der Behandlung von Whitespace deutlich. Ob und welche Art von Whitespace zwischen zwei Token im ursprünglichen Text steht, wird nicht repräsentiert. Wenn dort irgendetwas ungewöhnliches auftritt, wie z. B. abgerückte Satzzeichen oder doppelte Leerzeichen, dann kann der Text nicht von den Token rekonstruiert werden.

Man könnte argumentieren, dass die Dokumentaufbereitung Abweichungen von den „normalen Regeln“ der Typografie korrigieren, also z. B. Satzzeichen an die vorangehenden Wörter heranrücken müsse. Dies würde aber bedeuten, dass die Aufbereitung viele Aufgaben der Tokenisierung übernehmen müsste.

3.2.3 POS-Tagging und Lammatisierung

Beim Tagging wird jedes Token mit Informationen angereicht. Die Art der Informationen kann sehr unterschiedlich sein. Ebenso vielfältig sind die Anwendungen, bei denen Tagging nützlich ist. Einen Einblick bieten Leech und Smith (1999). Die Bezeichnung „Tag“, die mit „Etikett“ oder „Anhängsel“ übersetzt werden kann, deutet darauf hin, dass Tags sich immer auf genau ein Token beziehen. Der Aufbau tokenübergreifender Strukturen, wie z. B. beim Parsing, wird nicht unter Tagging zusammengefasst. Prinzipiell ist es aber möglich, Relationen zwischen Token mit Tags zu annotieren.

Im KoKS-System werden die Wortart (Part of Speech, POS) und das Lemma (die Grundform) jedes Tokens annotiert. Dazu wird der IMS TreeTagger[9] eingesetzt, der die Sprachen Deutsch und Englisch, die im KoKS-Projekt auftreten, unterstützt.[10]

Tagsets

Ein Tagset ist die Menge der Tags, die annotiert werden können. Der IMS TreeTagger verwendet für die unterstützen Sprachen unterschiedliche POS-Tagsets. Für Englisch ist es das Penn-Treebank[11] Tagset, für Deutsch das kleine (s.u.) STTS Tagset. Informationen zu den Tagsets stehen auf der Webseite zum IMS TreeTagger (siehe Fußnote 9) und zur Verfügung, die auch im KoKS-Abschlussbericht zusammengefasst sind.

Die Tagsets gehen über die Hauptwortarten deutlich hinaus. Sie umfassen 48 (Penn-Treebank) bzw. 54 (IMS TreeTagger) POS-Tags. Das STTS Tagset ist hierarchisch aufgebaut. Jedes Tag gehört zu einer von elf Hauptwortarten (Nomina, Verben, Artikel, Adjektive usw.) oder ist ein spezielles Tag, z. B. für Satzzeichen. Sieben Hauptwortarten sind weiter unterteilt in Unterwortarten. Beispielsweise sind Nomina gegliedert in Eigennamen und „normale Nomina“ (Zitat STTS Tagging Guideline[12]). Die Pronomina sind noch in einer dritten Hierarchieebene unterteilt. Das große STTS Tagset[13] gliedert die Tags noch

[9] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[10] Inzwischen stehen auch angepasste Versionen für Altfranzösisch, Französisch und Italienisch zur Verfügung.
[11] http://www.cis.upenn.edu/~treebank/
[12] Auf der TreeTagger Webseite verfügbar, siehe Fußnote 9.
[13] http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.ps.gz

Token	POS-Tag	Lemmata	Token	POS-Tag	Lemmata
Die	ART	d	The	DT	the
Fete	NN	Fete	school	NN	school
zum	APPRART	zum	’s	VBZ	be
Ferienbeginn	NN	Ferienbeginn	out	IN	out
fiel	VVFIN	fallen	party	NN	party
ins	APPRART	ins	was	VBD	be
Wasser	NN	Wasser	called	VBN	call
,	$,	,	off	RP	off
weil	KOUS	weil	.	SATZ-P	.
die	ART	d	<SATZ>		
Disco	NN	Disco	<segmentgrenze>		
abgebrannt	VVPP	abbrennen	The	DT	the
war	VAFIN	sein	club	NN	club
.	SATZ-P	.	had	VBD	have
<SATZ>			burned	VBN	burn
<segmentgrenze>			down	RP	down
Außerdem	ADV	außerdem	.	SATZ-P	.
kam	VVFIN	kommen	<SATZ>		
auch	ADV	auch	<segmentgrenze>		

Abbildung 3.3: getaggtes Dokumentpaar

weiter, um detailiertere Informationen, z. B. zu Kasus, Numerus und Genus bei Nomina, annotieren zu können, wird aber vom IMS TreeTagger nicht benutzt.

Die Lemmatisierung wird häufig nicht als Tagging, sondern nur als Nebenprodukt des POS-Tagging wahrgenommen, da sie von den meisten POS-Taggern optional angeboten wird. Da jedem Token eine Grundform zugewiesen wird, kann man auch hier von Tagging sprechen. Das Tagset umfasst alle Grundformen, die potentiell vom Tagger annotiert werden können. Im Falle des IMS TreeTaggers ist das Tagset eine endliche Menge, da die Grundformen mit der Vollform (also dem Token) nachgeschlagen werden.[14] Eine Analysekomponente, die unbekannte Wörter auf eine Grundform reduzieren kann, wird in der Beschreibung des IMS TreeTaggers nicht erwähnt.

Abbildung 3.3 zeigt einen Ausschnitt aus dem getaggten Beispieldokumentpaar. An den POS-Tags der Artikel beider Sprachen kann man auch ohne Kenntnis der Tagsets erkennen, dass sie unterschiedlich sind.

Funktionsweise

Viele POS-Tagger arbeiten laut van Halteren und Voutilainen (1999) in drei Schritten (Seite 109–110): Tokenisierung (bereits im Abschnitt 3.2.2 behandelt), Ermittlung der Tags, die für jedes einzelne Token in Frage kommen, und Auswahl eines Tags je Token mit Hilfe eines Modells der Sprache.

[14]Wenn mehrere Grundformen in Frage kommen, annotiert der IMS TreeTagger eine Liste von Grundformen. Da keine Grundform mehrmals aufgelistet wird, ist auch die Anzahl dieser Grundformlisten, die auch zum Tagset gezählt werden müssen, endlich.

Der IMS TreeTagger benutzt eine Vollformliste, um ein Token auf einen Wahrscheinlichkeitsvektor abzubilden. Das heißt, dass nicht nur aufgelistet wird, welche Tags für das betreffende Token möglich sind, sondern darüber hinaus auch eine Wahrscheinlichkeit für jedes POS-Tag angegeben wird. Ist das Token nicht verzeichnet, dann stehen dem IMS TreeTagger noch andere Methoden zur Verfügung, um zu einem Wahrscheinlichkeitsvektor zu gelangen, siehe Schmid (1994) und Schmid (1995). Beispielsweise können aus den letzten Zeichen des unbekannten Wortes Informationen gewonnen werden.

Im nächsten Schritt (dem dritten nach der Liste von van Halteren und Voutilainen) entscheidet der IMS Tagger welches Tag dem Token tatsächlich zugewiesen wird. Der Tagger nutzt wie viele andere POS-Tagger auch ein Markov Modell, innerhalb dessen mit dem Viterbi Algorithmus die wahrscheinlichste Tagsequenz gefunden wird. Die Übergangswahrscheinlichkeiten zwischen den Zuständen des Modells werden vorab aus einem Trainingskorpus, das manuell annotiert wurde, ermittelt. Hier wendet der IMS TreeTagger einen Decision Tree (Entscheidungsbaum) an, um Zustände zusammenzulegen. Auf diese Weise wird das so genannte Sparse Data Problem umgangen, das darin besteht, dass nicht genug Daten vorhanden sind, um alle Übergangswahrscheinlichkeiten zuverlässig abschätzen zu können. Der Entscheidungsbaum spielt also nur in der Trainingsphase eine Rolle. Das eigentliche Tagging bedient sich dann des Markov Modells, dessen Parameter im Training bestimmt wurden. Auf verschiedene Erweiterungen, die für das Training des deutschen Taggers notwendig waren, da dort das Trainingskorpus kleiner war, geht Schmid (1995) im zweiten Artikel ein.

Zum Verständnis der Artikel von Schmid sollte man mit verschiedenen bedingten Wahrscheinlichkeiten von Wort- und Tagsequenzen umgehen können. Eine gute Einführung bieten Manning und Schütze (1999) in einem Kapitel über Markov Modelle (Seite 318–340). Das anschliessende Kapitel über POS-Tagging (Seite 341–381) ist zur Vertiefung sicherlich lesenswert, aber zum Erarbeiten der genannten Artikel über den IMS TreeTagger nicht erforderlich. Weitere Bemerkungen zur Feinabstimmung des Markov Modells finden sich in Brants (2000). Toutanova et al. (2003) erläutern am Beispiel „will to fight“ das Unvermögen von POS-Taggern, die auf einem Markov Modell basieren, Informationen von Vorgängertoken und Nachfolgertoken gleichermaßen zu nutzen. Dies führe dazu, dass im Beispiel entweder „will“ als Verb oder „fight“ als Nomen getaggt wird.

Lemmatisierung

Zur Lemmatisierung erwähnt Schmid (1995) lediglich, dass beim Aufbau des Vollformlexikons, das die Wahrscheinlichkeitsvektoren der einzelnen POS-Tags aufnimmt, auch die Analyseergebnisse der Morphologiekomponente „DMOR“ einflossen (Abschnitt 4 „Tests“). Wie genau die Lemmatisierung funktioniert, kann den Quellen nicht entnommen werden. Vermutlich wurden auch die bei der DMOR-Analyse bestimmten Grundformen in das Vollformlexikon aufgenommen, sodass der Tagger in der Lage ist, diese zu annotieren. Die Lemmatisierung spielt in der Darstellung des POS-Taggers keine Rolle, ist also kein Nebenprodukt, sondern eine zusätzliche Leistung des IMS TreeTaggers.[15]

Wichtig für diese Arbeit (und auch für das KoKS-System) ist die Tatsache, dass der IMS TreeTagger keine Disambiguierung der Lemmata vornimmt. Kommen für ein Token mehrere Grundformen in Frage, dann annotiert der Tagger alle Alternativen. Tabelle 3.2 zeigt einige Beispiele aus dem Teilkorpus EU/1998. Die POS-Tags sind mit angegeben,

[15] Zumindest wird von der Möglichkeit, den Parameterraum des Markov Modells zu vergrößern, indem die Grundformen in die Zustände mit aufgenommen werden, und es dann dem Decision Tree Verfahren zu überlassen, den Parameterraum wieder geeignet zu verengen, keinen Gebrauch gemacht. Die entsprechenden bedingten Wahrscheinlichkeiten enthalten nur POS-Tags als Vorbedingung (siehe Formeln in Schmid (1995) Seite 2).

Häufigkeit	Token	POS-Tag	Lemmata
2	Andreas	NE	Andrea, Andreas
12	Antworten	NN	Antwort, Antworten
14	führen	VVFIN	fahren, führen
26	gelangt	VVPP	gelangen, langen
54	gewährt	VVPP	gewähren, währen
14	Listen	NN	List, Liste, Listen
15	Mitteln	NN	Mittel, Mitteln
23	Studien	NN	Studie, Studium

Tabelle 3.2: Token mit mehreren annotierten Grundformen (Auswahl)

Token	POS-Tag	Lemmata	Token	POS-Tag	Lemmata
Gefallen	NN	Gefallen	Gefallen	VVPP	fallen, gefallen
findet	VVFIN	finden	ist	VAFIN	sein
er	PPER	er	er	PPER	er
daran	PAV	daran	nicht	PTKNEG	nicht
bestimmt	VVPP	bestimmen	.	$.	.
.	$.	.	Gefallen	VVINF	gefallen
Gefallen	VVINF (*)	gefallen	wird	VAFIN	werden
wird	VAFIN	werden	es	PPER	es
sie	PPER	sie	ihr	PPOSAT (*)	ihr
nicht	PTKNEG	nicht	jedoch	ADV	jedoch
daran	PAV	daran	bestimmt	VVIMP	bestimmen
finden	VVINF	finden	nicht	PTKNEG	nicht
.	$.	.	.	$.	.

Abbildung 3.4: Einfluss der POS-Wahl auf die Lemmatisierung

da der IMS TreeTagger scheinbar die Liste der Grundformen auf solche Grundformen beschränkt, die mit dem für das Token bestimmte POS-Tag vereinbar sind. Ein geeignetes Token für einen Test des Verhaltens des Taggers ist „Gefallen“. In einem Kontext, in dem es als Nomen getaggt wird aber auch isoliert betrachtet ein Verb sein könnte, d. h. am Satzanfang steht, müssten auch die Verben „fallen“ und „gefallen“ annotiert werden, wenn das POS-Tag keine Rolle spielt. Abbildung 3.4 zeigt, dass je nach POS-Tag eine andere Grundformenliste annotiert wird. In den Testsätzen sind zwei POS-Taggingfehler enthalten, die in der Abbildung mit Sternchen markiert wird.

Im Deutschen sind viele Verben und Nomen betroffen. Im Englischen treten lexikalische Mehrdeutigkeiten innerhalb einer Wortklasse viel seltener, im gesamten KoKS-Korpus gar nicht, auf. Ein Beispiel wäre „saw“: Als Verb kann es die Vergangenheitsform von „see“ (sehen) und Präsenz von „saw“ (sägen) sein. (Des Weiteren kann es das Nomen „saw“ (Säge) sein.)

Schließlich muss bei den annotierten Grundformen beachtet werden, dass der IMS TreeTagger nicht alle Token, die in einer Eingabe auftreten können, in seiner Vollformenliste verzeichnet haben kann. Unbekannte Wörter erhalten die Grundform „<unknown>“.

Deutsch		Englisch	
Häufigkeit	Token	Häufigkeit	Token
7562	Mio.	5940	EU
4913	*	5219	ECU
4172	dass	3004	SPD
3251	EU	2398		
2749	Ron	2096	Hermione
2096			1648	DM
2002	Hermine	1169	Hagrid
1903	muss	1063	MECU
1480	Euro	1058	Dumbledore
1130	dich	942	Bundestag
1005	Hagrid	920	FDP
1000	Dumbledore	871	euro

Tabelle 3.3: Häufige Token mit unbekannter Grundform

Tabelle 3.3 zeigt die häufigsten betroffenen Token im KoKS-Korpus.

Fehlerrate

Wichtig für die Andwendungen in KoKS und in dieser Arbeit ist auch die Fehlerrate des Taggers. Der getaggte Text in Abbildung 3.4 offenbart bereits, dass der Tagger gelegentlich Fehler macht. Laut Schmid (1995) erreicht der POS-Tagger für das Deutsche 97,5 % und für das Englische 96,8 % Korrektheit. Da diese Zahlen auf einzelne Token bezogen sind, bedeutet dies trotz der hohen Korrektheit, dass sehr viele Sätze Fehler enthalten.

Für das KoKS-System ist die Fehlerrate niedrig genug. Tag-Sequenzen mit einer Länge von bis zu sechs Token sollten häufig korrekt sein, eine zufällige Verteilung der Fehler vorausgesetzt. Bei einer Translation Memory Anwendung, die auch POS-Tags für das Matching ganzer Sätze nutzt, können die Fehler jedoch Auswirkungen haben. Das wird im Kapitel 4 zu berücksichtigen sein.

3.2.4 Segmentierung

Unter Segmentierung versteht man die Zerlegung eines Textes in eine Sequenz von Segmenten. Die Art und Größe der Segmente kann je nach Zielsetzung sehr verschieden sein. In der Diskursanalyse werden sowohl grobe Segmentierungen, die vergleichbar sind mit der typographisches Dokumentstruktur (Abschnitte und Absätze), als auch sehr feine Segmentierungen, deren Segmente nur wenige Sätze umfassen, vorgenommen, siehe z. B. Sardinha (1997) Seite 5–8.

Im KoKS-System wird der Begriff Segment anders verstanden. In der Regel sind hier Segmente identisch mit Sätzen. Neben Satzgrenzen sind auch die während der Aufbereitung (siehe Abschnitt 3.2.1) eingefügten Absatzgrenzen Segmentgrenzen, sodass auch Überschriften ein Segment bilden. Segmente können aber im KoKS-System auch mehrere Sätze umfassen oder leer sein. Der Aligner (siehe Abschnitt 3.2.5) verschmilzt Segmente,

... den Zettel , der am ramponierten alten Notizbrett aufgetaucht war .
<segmentgrenze>
” Ende Oktober , an Halloween .
<segmentgrenze>
” ” Klasse ” sagte Fred, der Harry durch das Porträtloch gefolgt war , ”
ich muss zu Zonko , meine Stinkkügelchen sind fast alle .
<segmentgrenze>
” Harry ließ sich in den Sessel neben Ron fallen ; ...

Abbildung 3.5: Segmentierungsfehler bei wörtlicher Rede

um das Alignment zu repräsentieren. Nach dem Alignen besteht jedes Alignment-Bead aus genau einem deutschen und einem englischen Segment. In Abbildung 3.3 sieht man, wie vor dem Alignment jedes Satzende mit einem Segmentende zusammenfällt. Die Segmentendemarkierungen[16] nach dem Alignen zeigt Abbildung 3.6.

Segmente spielen im KoKS-System und im Translation Memory dieser Arbeit eine zentrale Rolle. Die Suche im Korpus erfolgt grundsätzlich segmentweise. Alle Indizes (siehe Abschnitt 3.2.7) verweisen auf Segment-Nummern. Eine gute Erkennung der Satzgrenzen ist daher sehr wichtig. Der IMS TreeTagger entscheidet bereits im Tokenisierungsmodul für jeden Punkt, ob er ein Satzende kennzeichnet. Ein Punkt wird nur als eigenständiges Token behandelt, wenn er als Satzzeichen eingestuft wurde. Der Tokenisierer verfügt über eine Abkürzungsliste und erkennt auch Fälle wie z. B. „der 5. Punkt der Tagesordnung“.

Die Qualität der Klassifizierung der Punkte konnte im KoKS-Projekt mit einfachen Regeln noch weiter erhöht werden. Z. B. wird grundsätzlich ein Satzende angenommen, wenn nach einem Punkt ein Wort groß geschrieben wird, dessen Lemma klein geschrieben wird. (Das Lemma wird vom IMS Tagger annotiert.) Details finden sich im Anhang des KoKS-Abschlussberichts.

Bei „Water Rats“ in Anführungszeichen (siehe Tabelle 3.1) verschluckt der KoKS-Satzendenerkenner das Wort „Rats“. Ist das Anführungszeichen nicht das letzte Zeichen der Eingabe, dann verschwindet dieser Fehler.

Wörtliche Rede

Ein spezielles Problem für die Segmentierung stellt wörtliche Rede dar. Da die Grenzen im KoKS-System nach Satzzeichen gezogen werden, gehört das schließende Anführungszeichen zum nächsten Segment, das dann je nach Situation eine ungerade Anzahl von Anführungszeichen enthält, mit zwei Anführungszeichen beginnt und/oder dessen Passagen genau invers in wörtliche Rede und normalen Text eingeteilt sind. Abbildung 3.5 zeigt einen kurzen Ausschnitt aus dem Harry-Potter Korpus, in dem einige dieser Probleme auftreten. Eine einfache Lösung, die aus Zeitgründen nicht mehr umgesetzt wurde, wäre, die Segmentendemarkierungen immer dann hinter ein Anführungszeichen zu verschieben, wenn die Anzahl der Anführungszeichen im aktuellen Segment ungerade ist. Pa-

[16] Die für die Markierung verwendete Zeichenfolge `<segmentgrenze>` ist irreführend. Dem letzten Segment muss eine Segmentmarkierung folgen, während vor dem ersten Segment keine Markierung stehen darf. Es handelt sich also um eine Endemarkierung und nicht um eine Grenzmarkierung.

trick Tschorn[17] weist darauf hin, dass dies Verfahren scheitert, wenn Anführungszeichen fehlen, z. B. aufgrund von OCR-Fehlern.

Ein weiteres Problem der wörtliches Rede ist ihre Verzahnung mit dem sie umgebenen Text. Mit einer sequentiellen Segmentierung können solche Sätze nicht getrennt werden. Aber selbst diskontinuierliche Segmente reichen zur Lösung des Problems nicht aus, da Satzzeichen von beiden verzahnten Sätzen gemeinsam genutzt werden können. Regelmäßig geschieht dies mit satzbeendender Interpunktion, wenn die Satzenden zusammenfallen. Abbildung 3.5 zeigt, dass auch Satzzeichen an anderen Stellen fehlen können. Wie soll maschinell entschieden werden, ob und welches Satzzeichen zwischen „Klasse“ und „ich muss zu Zonko“ gesetzt werden soll, wenn die wörtliche Rede als eigenes Segment extrahiert werden würde?

OSCAR Standard für Segmentierungsregeln

Die OSCAR Segmentation Working Group[18] hat einen ersten Entwurf für einen Standard zum Beschreiben von Segmentierungsregeln vorgestellt. Der Entwurf wird hier diskutiert, um zu zeigen, mit welchen Methoden die Arbeitsgruppenmitglieder glauben, die Segmentierung leisten zu können.[19] Grundlage des Standards sind reguläre Ausdrücke, die die möglichen Zeichenfolgen vor und nach einer Segmentgrenze angeben. Zusätzlich können Ausnahmeregeln formuliert werden. Wie die Regeln zusammenwirken, wird in dem Entwurf (noch) nicht definiert. Es sind keine Elemente vorhanden, um Ausnahmen auf bestimmte Regeln zu beziehen. Dies und das enthaltene Beispiel legen nahe, dass eine potenzielle Segmentgrenze mit linkem Konkext c_l und rechtem Kontext c_r genau dann als Segmentgrenze akzeptiert wird, wenn

$$\bigvee_{(b,a)\in R} (M(b+'\$', c_l) \wedge M('\hat{}'+a, c_r)) \wedge \neg \bigvee_{(b,a)\in E} (M(b+'\$', c_l) \wedge M('\hat{}'+a, c_r))$$

wahr ist, wobei M das Matching-Prädikat für reguläre Ausdrücke ist.

Im Entwurf wird ein Beispielregelsatz angegeben, der Punkt, Fragezeichen und Ausrufezeichen als Satzendemarkierungen erkennt und als Ausnahmen Nummerierungen von Listen und die Abkürzung „etc“ berücksichtigt. Um die XML-Notation des Standard hier nicht vorstellen zu müssen, sind die Regeln im folgenden als Mengen passend zur obigen Formel wiedergegeben:

```
R = {
   ('[\.\?!]+', '\s'),
   ('', '\n'),
}
E = {
   ('^\s*[0-9]+\.', '\s'),
   ('[Ee][Tt][Cc]\.', ''),
}
```

Die Muster `\s` und `\n` stehen dabei für beliebigen Leeraum und Zeilenumbrüche. (Letzteres Muster wurde im Entwurf nicht definieren.) Die erste Ausnahmeregel setzt voraus,

[17] persönliche Kommunikation

[18] `http://lisa.org/oscar/seg/`

[19] Dies muss nicht den Stand der Entwicklung in den einzelnen beteiligten Unternehmen wiederspiegeln. Womöglich werden Informationen zurückgehalten, um den Mitbewerbern keinen Einblick in die eigenen Lösungsansätze zu gewähren. Dann wäre allerdings zu Fragen, warum überhaupt an einem gemeinsamen Standard für den Austausch von Segmentierungsregeln gearbeitet wird.

Token	POS-Tag	Lemmata	Token	POS-Tag	Lemmata
Die	ART	d	The	DT	the
Fete	NN	Fete	school	NN	school
zum	APPRART	zum	’s	VBZ	be
Ferienbeginn	NN	Ferienbeginn	out	IN	out
fiel	VVFIN	fallen	party	NN	party
ins	APPRART	ins	was	VBD	be
Wasser	NN	Wasser	called	VBN	call
,	$,	,	off	RP	off
weil	KOUS	weil	.	SATZ-P	.
die	ART	d	<SATZ>		
Disco	NN	Disco	The	DT	the
abgebrannt	VVPP	abbrennen	club	NN	club
war	VAFIN	sein	had	VBD	have
.	SATZ-P	.	burned	VBN	burn
<SATZ>			down	RP	down
<segmentgrenze>			.	SATZ-P	.
Außerdem	ADV	außerdem	<SATZ>		
kam	VVFIN	kommen	<segmentgrenze>		
auch	ADV	auch	The	DT	the

Abbildung 3.6: aligntes Dokumentpaar

dass vorangehende Segmentgrenzen bereits identifiziert wurden, die Abarbeitung also sequentiell erfolgt.

Es wird nicht angedeutet, ob Zugriffsmöglichkeiten auf ein Lexikon geplant sind. Dies wäre sinnvoll, um nicht für jede einzelne Abkürzung eine Regel formulieren zu müssen. Ebensowenig kann mit tokenisiertem Text umgegangen werden, da keine Muster für Tokengrenzen definiert werden. POS-Muster und Grundformen können nicht für die Segmentierung herangezogen werden.

3.2.5 Alignment

Eine abstrakte Darstellung des Alignments wurde bereits in Abschnitt 2.2.3 vorgenommen. Da die Betrachtung dort sehr allgemein ist, wurde die KoKS-Terminologie nicht übernommen. Da die Einheiten, die alignt werden, überwiegend Sätze sind, wird im folgenden vereinfachend von Sätzen gesprochen, obwohl auch Überschriften und Listenelemente Einheiten sein können. In KoKS heißen die Gruppen eines Alignment-Beads Segmente, wie bereits im Abschnitt 3.2.4 erwähnt wurde. Leider wird die Segmentendemarkierung auch verwendet, um die Einheiten zu kennzeichnen, aus denen der Aligner die Gruppen bilden darf, sodass mit Segment auch eine einzelne Einheit gemeint sein kann. Der Unterschied zwischen Abbildung 3.3 und 3.6 zeigt, wie die Markierungen verändert werden, um das Alignment zu repräsentieren. (In dem abgebildeten Ausschnitt liegt ein 1 : 2 Alignment-Bead vor.)

Der KoKS-Aligner ist auf Satzalignment spezialisiert. Gruppen können nur aus zusammenhängenden Einheiten gebildet werden, und die Zuordnungen dürfen sich nicht

überkreuzen. Etwas ungewöhnlich für einen Satzaligner ist, dass der KoKS-Aligner zwar keine leeren Gruppen erlaubt, aber zugleich die Anzahl der Einheiten in einer Gruppe nicht nach oben beschränkt. Ein KoKS-Alignment ist also eine Abfolge von $n : m$ Zuordnungen mit $n, m > 0$.

Die Beschreibung des Aligners ist im KoKS-Abschlussbericht bereits sehr ausführlich. Hier wird trotzdem auf die Funktionsweise eingegangen, da das Alignment der Schlüssel zur Identifikation der Übersetzung innerhalb eines Translation Memorys ist. Des Weiteren wird hier eine andere Sichtweise auf den KoKS-Aligner vorgestellt, mit der die konzeptionellen Defizite des KoKS-Aligners besser verstanden werden können und aus denen sich Verbesserungsmöglichkeiten ableiten lassen.[20]

Abstandswerte und -matrix

Der KoKS-Aligner bestimmt nicht direkt die Abstände von Gruppen der beiden Sprachseiten Deutsch und Englisch. Es werden immer nur einzelne Sätze miteinander verglichen. Das hat den Vorteil, dass nicht so viele Kombinationen von zu vergleichenden Satzgruppen auftreten. Wenn das deutsche Eingabedokument m Sätze und das englische n Sätze umfasst, dann müssen maximal mn Abstandswerte berechnet werden. Diese Werte können vorab bestimmt und in einer Matrix, die Abstandsmatrix, abgelegt werden, auf die der Alignment-Optimierer zurückgreift.[21]

In die Berechnung der Abstandswerte fließen verschiedene, linguistisch motivierte Bewertungen ein. Es werden die POS-Tags und Lemmata genutzt, die vom IMS TreeTagger annotiert wurden, und auf ein umfangreiches, bilinguales Wörterbuch zurückgegriffen, das im KoKS-Projekt aus verschiendenen Quellen zusammengestellt wurde.

Zu Wörtern aus offenen Wortklassen werden die Entsprechungen zwischen den Sätzen gezählt, die mit Hilfe des KoKS-Wörterbuchs und den annotierten Grundformen gefunden werden können. Die übrigen Wörter aus offenen Wortklassen werden zu einer Zeichenkette je Sprachseite zusammengefügt und mit einem Abstandsmaß verglichen, das bereits auf kurze übereinstimmenden Zeichenfolgen anspricht und die Reihenfolge der Übereinstimmungen nachrangig behandelt. Schließlich werden die Wörter aus geschlossenen Wortklassen gezählt, um ihre Anzahl zu vergleichen. Weitere Informationen, z. B. der Anteil der einzelnen Wortarten, werden nicht ausgewertet.

Da die Abstandswertberechnung viel Zeit beansprucht, werden unter verschiedenen Bedingungen Werte durch den minimalen oder maximalen Abstandswert abgeschätzt. Betroffen sind hiervon beispielsweise Sätze aus Absätzen, die sich nicht entsprechen. (Siehe KoKS-Abschlussbericht für Details.) Das Laufzeitverhalten des KoKS-Aligners ist trotzdem mindestens quadratisch, da die volle Abstandsmatrix mit mn Einträgen erzeugt werden muss und die Dokumentlängen m und n deutlich korrelieren.[22] In der Praxis ist vor allem ein Problem, dass der Speicherbedarf der Abstandsmatrix quadratisch mit der Länge der Eingabedateien wächst.

[20]In diesem Zusammenhang möchte der Autor auch Patrick Tschorn, der wesentlich Komponenten des KoKS-Aligner entwickelt hat, für die zahlreichen Gespräche über Alignment danken.

[21]Ausschlaggebend für diese Trennung war im KoKS-Projekt, dass so die Entwicklung des Aligners auf zwei Projektmitglieder verteilt werden konnte. Später (nach der Einführung der Umlautkorrektur) konnten gespeicherte Abstandsmatrizen tatsächlich wiederverwertet und so mehrere Tage Rechenzeit eingespart werden.

[22]Im KoKS-Projekt wurden zwar einige Komponenten für eine kompaktere Repräsentation der Matrizen angepasst. Es gelang aber nicht mehr, ein reibungsfreies Zusammenspiel herzustellen, sodass auf eine Darstellung, die sämtliche Werte der Matrix auflistet, nicht ganz verzichtet werden konnte.

$i \backslash j$	1	2	3	4	5	6	7	8	9
1	**1**	1	1	1	1	1	1	1	1
2	1	**3**	5	7	9	11	13	15	17
3	1	5	**13**	25	41	61	85	113	145
4	1	7	25	**63**	129	231	377	575	833
5	1	9	41	129	**321**	681	1 289	2 241	3 649
6	1	11	61	231	681	**1 683**	3 653	7 183	13 073
7	1	13	85	377	1 289	3 653	**8 989**	19 825	40 081
8	1	15	113	575	2 241	7 183	19 825	**48 639**	108 545
9	1	17	145	833	3 649	13 073	40 081	108 545	**265 729**
10	1	19	181	1 159	5 641	22 363	75 517	224 143	598 417
11	1	21	221	1 561	8 361	36 365	134 245	433 905	1 256 465
12	1	23	265	2 047	11 969	56 695	227 305	795 455	2 485 825
13	1	25	313	2 625	16 641	85 305	369 305	1 392 065	4 673 345
14	1	27	365	3 303	22 569	124 515	579 125	2 340 495	8 405 905
15	1	29	421	4 089	29 961	177 045	880 685	3 800 305	14 546 705
16	1	31	481	4 991	39 041	246 047	1 303 777	5 984 767	24 331 777

Abbildung 3.7: Anzahl der Pfade in der Abstandsmatrix

Pfadrepräsentation eines Alignments

In einer Abstandsmatrix fallen in der Regel längere Diagonalfolgen von Matrixzellen mit niedrigen Abstandswerten auf. Sie deuten auf Sequenzen von 1 : 1 zu alignenden Sätzen hin. Im KoKS-Projekt wurde daher entschieden, zum Bestimmen eines Alignments einen Pfad in der Abstandsmatrix zu suchen, der über Zellen führt, deren Abstandswerte in der Summe möglichst klein sind. Der Pfad soll die Zellen $(1,1)$ und (m,n) verbinden, da angenommen wird, dass das erste Alignment-Bead mindestens die ersten Sätze der zu alignenden Dokumente und entsprechend das letzte Bead die letzten Sätze enthält.

Jeder Pfad setzt sich aus einer Abfolge von Zellen zusammen. Nachfolger einer Zelle (i,j) können $(i+1,j)$, $(i,j+1)$ und $(i+1,j+1)$ sein, sofern sie innerhalb der Matrix liegen. Graphentheoretisch gesprochen handelt es sich um einen gerichteten Graphen mit mn Knoten und $(m-1)(n-1)+n(m-1)+m(n-1) = 3mn-2(m+n)+1$ Kanten. Die Zahl der möglichen Pfade von $(1,1)$ zu jeder einzelnen Zelle zeigt Abbildung 3.7 für eine 9 x 16 Matrix. In der Darstellung liegt $(1,1)$ oben links. In dieser Matrix kann man die Anzahl der möglichen Alignmentpfade für verschieden große Abstandsmatrizen ablesen. Beispielsweise gibt es 41 Alignmentpfade in einer 5 x 3 Abstandsatrix. Eine einfache, nicht rekursive Formel für die Anzahl der Pfade liegt nicht nahe. Im KoKS-Abschlussbericht wird ein exponentielles Verhalten zur Größe der Matrix vermutet. Die Werte in der Nähe der in der Abbildung hervorgehobenen Diagonalen wachsen überexponentiell zu $i+j-2$.[23]

Wie ein Pfad als Alignment interpretiert werden kann, ist nicht offensichtlich. Andere

[23] Bei einer Beschreibung der Pfadanzahl v mittels $v = b(i,j)^{i+j-2}$ liegen die Basen $b(i,j) = \sqrt[i+j-2]{v}$ in einem Bereich der Matrix über zwei, der sich ca. ±27 Grad um die Diagonale herum öffnet. Soweit die Folge $b(i,i)$ mit dem Python Modul „math" berechnet werden kann und vorausgesetzt, es treten keine numerischen Probleme auf, wächst sie streng monoton mit abnehmender Zuwachsrate. Die größte quadratische Matrix, die berechnet werden konnte, reicht bis $i = 405$. Die Basen wachsen über $2,4$ nur noch sehr langsam. Möglicherweise konvergiert die Folge, sodass die Pfadanzahl in $O(b^{i+j-2})$ mit $b > 2,403$ liegt.

```
                                    Xooooo  Xooooo
        Xooo   Xooo   Xoooo  oXXXXo  oXXXoo
Xoooo   oXXo   oXoo   oXooo  ooooXo  ooooXo
oXXXo   ooXo   oXXo   oXXXo  ooooXo  ooooXo
ooooX   oooX   oooX   ooooX  oooooX  oooooX

a) 3:1  b) 2:2  c) 2:2  d) 3:2  e) 4:3   f) 3:1 + 1:2
```

Abbildung 3.8: Pfadrepräsentation von Alignments

Zuordnungen als 1 : 1 Zuordnungen treten immer dann auf, wenn der Pfad nicht diagonal verläuft. Eine rechte oder untere Nachbarzelle vergrößert das aktuelle Alignment-Bead um die Sätze, deren Abstand die Matrixzelle enthält. Abbildung 3.8 zeigt einige Pfade und die Art der Zuordnung. Die einzelnen Zeichenpositionen entsprechen Zellen einer Abstandsmatrix. Die Zellen, über die der jeweilige Pfad führt, sind mit X markiert. Oben links und unten rechts in jedem Teilbild ist der weitere Verlauf des Pfades angedeutet. Teilbilder b und c zeigen, dass es für $m:n$ Zuordnungen mit $\min(m,n) > 1$ immer zwei mögliche Pfadeverläufe gibt. In e/f wird deutlich, dass kleine Änderungen zu einem ganz anderen Alignment führen können.

Optimierung

Der KoKS-Aligner sucht einen Pfad in der Abstandsmatrix mit möglichst geringer Summe der Abstandswerte. Die Suche wird mit dem A-Stern-Algorithmus und einer Heuristik, die die minimale Abstandssumme zwischen zwei beliebigen Matrixzellen abschätzt, effizient durchgeführt. So konnte selbst eine 699 x 685 Matrix in wenigen Minuten verarbeitet werden, obwohl die Anzahl der möglichen Pfade bei $6,6 * 10^{528}$ liegt.

Teilpfade wie in b bis e (Abbildung 3.7) können nur gewählt werden, wenn eine Abkürzung der Ecke wie in Teilbild f nicht zu einer geringeren Abstandswertsumme führt. Das ist nur möglich, wenn die Eckzelle den Abstandswert null hat, da negative Abstandswerte nicht erlaubt sind.[24] Treten k solche Eckzellen am Alignment-Pfad auf, dann gibt es 2^k optimale Pfade. Welchen der Alignmentoptimierer wählt, hängt von Details der Implementation ab. Da nicht positive Abstandswerte sehr ungewöhnlich sind, erzeugt der KoKS-Aligner also im Regelfall nur $n:1$ und $1:n$ Zuordnungen mit $n \geq 1$. [25]

Ausblick

Um die hier geschildertert Probleme des Aligners und andere zu lösen, die bereits im KoKS-Abschlussbericht beschrieben werden, wurde eine neue Pfadrepräsentation und Pfadbewertung entworfen und implementiert. Die Repräsentation erlaubt alle Zuordnungsarten,

[24] Das KoKS-Abstandsmaß gibt leider doch negative Werte aus. In den vorhandenen Abstandsmatrizen wurden Werte zwischen -10^{-8} und -10^{-9} beobachtet. Vermutlich sind numerische Probleme die Ursache und die Werte müssten eigentlich null sein.

[25] Es wurde nochmal der Quellcode des Aligners durchgesehen, ob nicht doch weitere Faktoren in die Pfadbewertung einfließen. Des Weiteren wurde mit einer manuell erstellten Matrix versucht, eine 3 : 3 Zuordnung zu erzwingen. Ebenso wurden die Alignmentpfade zu 10 mit Zufallswerten gefüllten 51 x 52 Matrizen bestimmt. Auch hier trat kein Pfad auf, der über Eck führt.

auch $n : 0$. Beibehalten wurde, dass die Gruppen zusammenhängend sein müssen und nicht über kreuz alignt werden können. Die Beschränkung der Abstandswertberechnung auf Satzpaare wurde aufgegeben zugunsten einer Berechnung nach Bedarf für beliebige Gruppenpaare. Erste Experimente zeigten ein gutes Laufzeitverhalten. Jedoch war keine Zeit vorhanden für einen gründlichen Test des Aligners und die Feinabstimmung der Parameter.

Es ist unklar, ob sich der Aufwand für die Entwicklung eines neuen Aligners lohnt, da der KoKS-Aligner bereits eine (für die Anwendungen im KoKS-Projekt und in dieser Arbeit) zufrieden stellende Alignmentqualität erreicht. Das ist ein weiterer Grund, warum der Ansatz nicht weiter verfolgt wurde.

3.2.6 Datenbank

Die alignten Dokumente werden in einer Datenbank abgelegt. Dabei wird jedes Tupel aus Token, POS-Tag, Grundform und Sprache nur einmal gespeichert und mit einer eindeutigen Zahl, der Token-ID, versehen, die zur Identifikation des Tupels dient. Der Dokumentinhalt reduziert sich damit auf eine Folge von Token-IDs, die in der Datenbank als funktionale Relation zwischen den natürlichen Zahlen und den Token-IDs realisiert ist. Als Tabelle dargestellt ergibt das die zwei Spalten Token-Nummer und Token-ID.

Das Alignment wird in der Datenbank repräsentiert, indem zu jeder Token-Nummer vermerkt wird, zu welchem Segment das Token gehört. Es werden gleiche Segmentnummern für die beiden Segmente eines Alignment-Beads verwendet. Entsprechend ist auch eine Satznummer vermerkt. (Gemeint ist die sprachliche Einheit Satz, nicht ein Datensatz.) Zu jedem Satz wird die Herkunft (Quelle, Autor und Jahr) und nochmal die Sprache gespeichert. Die Sprache wird aufgelistet, um ohne Rückgriff auf die Tokentupel Segmente einer bestimmten Sprache abfragen zu können. Eventuell war auch vorgesehen, dass Tokensprache und Satzsprache abweichen können. Die KoKS-Vorverarbeitung unterstützt dies jedoch nicht. Warum diese Informationen gerade bei Sätzen und nicht bei größeren Einheiten wie Absäztzen oder Dokumenten vermerkt werden, ist nicht (mehr) bekannt.

Einen weiteren Teil der Datenbank nehmen Indizes ein. Indizes auf Zeilenwerte einzelner Spalten und Kombinationen von Spalten werden von der Datenbanksoftware angeboten und automatisch und transparent bei SQL-Anfragen[26] eingesetzt. Darüber hinaus wurden spezielle Indizes aufgebaut, die eigene Tabellen erforden, beispielsweise eine Auflistung aller Segmentnummern sortiert nach Satzanfängen. Im nächsten Abschnitt 3.2.7 werden diese Indizes vorgestellt.

Behandlung der Grundformen

Das Tokentupel enthält die Grundform so, wie sie der Tagger annotiert. Bei manchen Token ist dies nicht eine einzelne Grundform, sondern eine Liste aus mehreren, durch senkrechte Striche getrennte Grundformen. Tabelle 3.2 im Abschnitt 3.2.2 zeigt ausgewählte Beispiele. Wenn nach Stellen im Korpus gesucht wird, die Token mit einer vorgegebenen Grundform enthalten, werden diese Grundformenlisten vom KoKS-System nicht berücksichtigt. Dies hat sowohl Vor- als auch Nachteile. Zum einen werden viele relevante Stellen mit Token, in deren Grundformenliste die gesuchte Grundform erscheint, nicht gefunden. Zum anderen werden falsche Treffer vermieden, die auftrete würden, wenn in einer Grundformenliste, die die gesuchte Grundform enthält, eine andere Grundform zutrifft. Im KoKS-System wurde also Wert darauf gelegt, dass möglichst viele Fundstellen korrekt sind, die

[26] SQL ist eine Sprache, in der eine Anwendung Fragen und Anweisungen an die Datenbanksoftware richtet.

Precision also hoch ist. Das geht auf Kosten des Recalls, also des Anteils der gefundenen (und korrekten) Fundstellen an den im Korpus tatsächlich vorhandenen, relevanten Stellen.

Im Rahmen dieser Magisterarbeit wurde eine zusätzliche Tabelle in der Datenbank angelegt, die die einzelnen Grundformen der Grundformenlisten verzeichnet und auf die jeweiligen Tokentupel verweist.[27] Es wurde ein Modul implementiert, dass zu einer Grundform alle infrage kommenden Token-Nummern ermittelt und darauf basierend verschiendene Suchmöglichkeiten im Korpus anbietet. Beispielsweise besteht die Möglichkeit, die Vollform in die Suche mit einzubeziehen. Dies kann sinnvoll sein, wenn die Grundform im System unbekannt ist. Der IMS TreeTagger annotiert als Grundform „<unknown>“, wenn ein Token nicht in seinem Vollformlexikon enthalten ist. Da es in dieser Arbeit darum geht, das Korpus als Informationsquelle für die Übersetzung zu nutzen und die Nützlichkeit abzuschätzen, ist ein hoher Recall wichtiger ist als gute Precision.

Eine alternative Lösung des Problems wäre die Disambiguierung der Grundformen. Denkbar wäre, einfache Regeln für die häufigsten Token von Hand zu erstellen. Beispielweise könnte man bei „führen“ heranziehen, ob „nach“ oder „zu“ in der Nähe auftritt. Wenn nur die häufigsten Token behandelt werden, ist der Aufwand nicht allzu hoch und trotzdem eine deutlich Verbesserung der Lemmatisierung möglich. Zu beachten ist, dass Regeln nicht jeden Fall, der in von Menschen verfassten Texten auftritt, berücksichtigen können. Eine Disambiguierung wird Fehler einführen, sodass im Vergleich zu der KoKS-Lösung die Precision der Anfrageergebnisse und im Vergleich zur neuen Lösung der Recall sinkt.

Detailprobleme

Im KoKS-Projekt konnte nicht jedes Detail der Implementation perfekt umgesetzt werden. Dafür fehlte die notwendige Zeit. So verwendet die SQL-Anfragesprache der Datenbank Anführungszeichen, um Werte, die selbst Zeichenfolgen sind, zu Kennzeichnen. In der KoKS-Implementation werden alle Anführungszeichen einfach in ein Nummernzeichen (#) verwandelt. Die bessere Lösung wäre gewesen, in der SQL-Dokumentation nachzuschauen, wie Anführungszeichen geschützt werden müssen, und eine entsprechende Funktion zu implementieren. In den im Rahmen dieser Magisterarbeit erstellten, neuen Softwarekomponenten wurde dies umgesetzt, da im Harry-Potter Korpus oft wörtlich Rede vorkommt. Die Umstellung sämtlicher Komponenten wurde aber aus Zeitmangel aufgegeben.

Die unvollständige Umstellung führt leider zu neuen Problemen. Eine Anfrage, die Anführungszeichen enthält, findet im Korpus keine Treffer. Erst eine Umstellung der gesamten Korpusvorverarbeitung würde hier Abhilfe schaffen. In dieser Arbeit tritt das Problem nicht auf, da für die Anfragen nur Sätze aus dem Korpus selbst verwendet werden.

3.2.7 Indizierung

Die Struktur der KoKS-Datenbank erlaubt einen sehr schnellen Zugriff auf alle Segmente, die ein bestimmtes Tokentupel (Token, POS-Tag, Grundform, Sprache) enthalten. Die Datenbank kann dabei auch Listen von Tokentupeln verarbeiten, von denen eines im Segment auftreten muss, damit das Segment gefunden wird. Auf diese Weise können alle Segmente zu z. B. einer Grundform und Sprache unabhängig von POS-Tag und Token mit einer Datenbank-Anweisung abgefragt werden.

Komplexere Anfragen bereiten jedoch Probleme. Beispielsweise möchte man alle Segmente erfragen können, die eine Kombination von Wörtern oder Grundformen enthalten.

[27] Im konkreten Datenbankdesign sind die Tokentupel auf mehrere Tabellen (Token, Grundformen, POS-Tagset) aufgeteilt. Der Verweis auf die Grundformenliste erfolgt über die Grundform-ID.

Im KoKS-Projekt wurde diese Anfrage umgesetzt, indem außerhalb der Datenbank die Segmentnummerlisten der einzelnen Wörter geschnitten werden. Dies ist keine gute Lösung, da die Einzellisten sehr lang sein können und deren Übertragung von der Datenbanksoftware zur Anwendung ineffizient ist. Eine vom Autor dieser Arbeit gefundenen Lösung, die innerhalb der Datenbank die Listen schneidet, läuft um ein Vielfaches, aber nicht um Größenordnungen schneller als die KoKS-Lösung.[28]

Die für die Anwendungen wichtigen Anfragen müssen also auf andere Weise beschleunigt werden. Im KoKS-Projekt, im Anschluss an den Projekt und im Rahmen dieser Arbeit wurden vom Autor verschiedene Indizes erstellt, die in Folgendem kurz vorgestellt werden.

Grundlagen

Die Zeilen einer Tabelle werden in einer Datenbank ungeordnet abgelegt, um die Datenhaltung möglichst einfach und anwendungsunabhängig zu halten.[29] Neue Zeilen können sehr schnell hinzugefügt werden, da nur der notwendige Platz geschaffen werden muss. Für Anwendungen, die hauptsächlich Informationen zusammentragen, beispielsweise Ereignisse protokollieren, kann dies wichtig sein. Würden die Zeilen sortiert gespeichert, müssten weitere Verwaltungsstrukturen für jede neue Zeile angepasst werden.

Sollen Zeilen mit vorgegebenen Spaltenwerten in einer unsortierten Tabelle ausgelesen, verändert oder gelöscht werden, muss die gesamte Tabelle durchsucht werden. Bei großen Tabellen kann dies sehr viel Zeit in Anspruch nehmen. Anwendung, die diese Operationen verwenden, würden also von zusätzlichen Datenstrukturen, die den Zugriff auf Zeilen mit vorgegebenen Spaltenwerten beschleunigen, profitieren. Indizes dienen genau diesem Zweck. Der Benutzer (oder der Verwalter der Datenbank) kann angeben, zu welchen Spalten oder Kombinationen von Spalten Strukturen aufgebaut und gepflegt werden sollen, die spätere Anfragen beschleunigen.

MySQL verwendet eine spezielle Baumstruktur, den B*-Baum, für Indizes. Diese Struktur erlaubt ein effizientes Suchen, Verändern, Einfügen und Löschen von Indexeinträgen. Blendet man den Aspekt der Effizienz aus, kann ein MySQL-Index als alphabetisch (oder numerisch) sortierte Liste aller Werte der indizierten Spalte mit einem Verweis auf die Zeilen, die den jeweiligen Wert aufweisen, verstanden werden.[30] Auf dieser Betrachtungsebene ist ein MySQL-Index wie ein Index eines Buches aufgebaut. Die Stichwörter entsprechen den Werten, die in der indizierten Spalte auftreten, und die angegebenen Seitenzahlen den Verweisen auf die Zeilen der Tabelle.

Die alphabetische Reihenfolge der Indexeinträge ermöglicht nicht nur ein schnelles Auffinden von Tabellenzeilen mit vorgegebenen Spaltenwerten. Auch Bereichsanfragen können mit solchen Indizes effizient ausgeführt werden. Wenn beispielsweise alle Zeilen mit Werten zwischen „Imperium“ und „Import“ gesucht werden, muss nur ein zusam-

[28]Realisiert ist dies über eine *n*-malige Verknüpfung der Korpustabelle mit sich selbst, wobei *n* die Anzahl der vorgegebenen Tokentupel ist, die im Segment auftreten sollen. Im KoKS-Projekt wurde davon ausgegangen, dass eine anwendungsseitige Lösung notwendig sei, vermutlich weil die von der eingesetzten Version der MySQL-Datenbanksoftware unterstützten Elemente der Abfragesprache SQL für unzureichend gehalten wurden. (Die Version unterstützt beispielsweise keine Subselects.)

[29]Der in MySQL verwendete Tabellentyp „MyISAM“ enthält zwar die Bezeichnung ISAM (index sequential access method, eine Methode, bei der die Daten sortiert abgelegt werden und ein dünn besetzter Index verwendet wird). MySQL setzt aber ohne Anweisung keine Indizes ein und erzeugt voll besetzte Indizes, wenn der Benutzer einen Index wünscht.

[30]MySQL unterstützt auch Indizes zu Kombinationen von Spalten. Die Sortierreihenfolge richtet sich dann nach der ersten in den Index einbezogenen Spalte. Bei gleichen Werten wird die nächste Spalte herangezogen. Typisches Beispiel ist die Kombination von den Spalten „Nachname“ und „Vorname“ in einer Tabelle mit Personendaten. Mehrdimensionale Suchbäume, z. B. *k*-d-Bäume, die beispielsweise für kartesische Koordinaten sinnvoll sind, werden von MySQL nicht unterstützt.

menhängender Bereich im Index gelesen werden.[31] Ebenso können alle Werte, die mit einem Präfix, z. B. „Imp", beginnen, schnell gefunden werden. Von dieser Möglichkeit wird bei den weiter unten beschriebenen Indizes Gebrauch gemacht.

Die Indizes einer Datenbank verhalten sich völlig transparent. Man muss nur einmal angeben, dass sie erstellt werden sollen, und schon verwendet die Datenbank sie automatisch, um die Bearbeitung von Anfragen zu beschleunigen. Für die im folgenden beschriebenen Indizes gilt dies nicht. Sie sind spezielle Tabellen, die zwar innerhalb der Datenbank gespeichert sind, aber explizit in einer SQL-Anweisung eingebunden werden müssen. Ebenso muss die Anwendungssoftware dafür Sorge tragen, dass diese Tabellen konsistent zum Korpus gehalten werden.[32] Das Nachschlagen innerhalb der Tabellen der manuellen Indizes erledigt die Datenbank wie für andere Tabelle auch über eigene Indizes.

Satzindex

Der einfachste, manuelle Index im KoKS-System listet alle Segmente auf. Im Regelfall sind dies Sätze, sodass hier vereinfachend von Sätzen gesprochen werden kann. Für jeden Satz werden die Token durch ein spezielles Zeichen getrennt zu einer Zeichenkette zusammengesetzt und zusammen mit der Segmentnummer in einer Tabelle aufgeführt. Um Speicherplatz zu sparen, wurden nur die ersten 56 Zeichen gespeichert. Die meisten Sätze können trotzdem eindeutig identifiziert werden. Um auch in den Fällen, in denen verschiedene Sätze mit der gleichen Wendung beginnen, eine möglichst kleine Treffermenge erhalten zu können, wird zusätzlich die Satzlänge in Token und die Sprache vermerkt.

Prinzipiell wären auch andere Eigenschaften der Sätze zum Einschränken der Treffermenge geeignet. Wenn die Eigenschaften so gewählt sind, dass unterschiedliche Sätze sehr selten die gleichen Eigenschaften haben, dann ist die Spalte, die die Satzanfänge enthält, zum Auffinden von Sätzen nicht nötig. Werden darüber hinaus die Eigenschaften auf den Wertebereich eines kurzen Datentyps der Datenbank abgebildet, dann belegt der Index sehr wenig Speicherplatz.

Abbildung 3.9 zeigt einen Ausschnitt aus der Tabelle zusammen mit einer SQL-Anfrage, die die Einträge von „Imperium" bis „Import" mit der Sprache „Deutsch" (kodiert mit dem Wert 1) auswählt und die Spaltennamen für die Ausgabe umbenennt.[33] Die Spalte für die Sprache wurde nicht abgebildet, da sie in den ausgewählten Zeilen nur den Wert 1 hat. Zwei Zeilen enthalten englischen Text. Dies ist weder ein Fehler des Moduls für die Indexerstellung noch der KoKS Datenbank. Die POS-Tags und Grundformen sind die, die sich einstellen, wenn der englische Text vom IMS TreeTagger für das Deutsche getaggt wird. Für das Segment 422412 hat eine Recherche in den beim Taggen erstellten Dateien ergeben, dass mindestens ein deutsches Dokument einen englischsprachigen Anhang enthält.

Das Auffinden eines Satzes erfolgt nun, indem er mit der gleichen Funktion wie bei der Erstellung des Indexes auf eine maximal 56 Zeichen lange Zeichenkette abgebildet und die Anzahl der Token bestimmt wird. Mit diesen Daten wird dann in der Index-Tabelle nachgeschlagen. Sofern die 56 Zeichen nicht den gesamten Anfragesatz abdecken, müssen die Sätze, auf die verwiesen wird, noch daraufhin überprüft werden, ob sie tatsächlich identisch mit dem Anfragesatz sind.

[31] Wenn die Blätter des B*-Baums nicht verkettet sind, dann stehen die Indexeinträge nicht explizit zusammen. Mit einer Traversierung des Baums startend mit dem Pfad zum ersten relevanten Eintrag und endend, sobald ein nicht relevanter Eintrag erreicht wird, kann der Indexbereich trotzdem effizient ermittelt werden.

[32] MySQL unterstützt keine Stored Procedures und Triggers.

[33] Es wurden anwendungsunabhängige Spaltennamen gewählt, da erwartet wurde, dass das Modul für diesen Index auch in anderen Zusammenhängen benutzt werden könnte, in denen die ganzzahligen Beschränkungen andere Bedeutungen haben.

```
mysql> SELECT name, beschr1 AS '#', datum AS 'SegNr.' FROM token_strict
    -> WHERE name BETWEEN 'Imperium' AND 'Import' AND beschr2 = 1;
+------------------------------------------------------+----+--------+
| name                                                 | #  | SegNr. |
+------------------------------------------------------+----+--------+
| Impfstoffe|gibt|es|bislang|nicht|.                   |  6 | 497752 |
| Impfstoffe|werden|nur|in|Notfällen|eingesetzt|.      |  7 | 446191 |
| Impfstoffe|werden|nur|in|Notsituationen|verwendet|.  |  7 | 456814 |
| Impfungen|gegen|Typhus|im|Hochwassergebiet          |  5 | 495533 |
| Impfungen|in|den|Niederlanden|Der|Ausschuss|gab|eine|bef | 39 | 466312 |
| Impfung|gegen|Typhus                                 |  3 | 566826 |
| Implementation|of|the|various|Directives|and|social|part | 19 | 422412 |
| Implementierung|und|Zusammenschaltung|europaweiter|Netze | 68 | 640588 |
| Implementing|the|euro|does|not|therefore|lead|to|any|del | 11 | 437346 |
+------------------------------------------------------+----+--------+
9 rows in set (0.00 sec)
```

Abbildung 3.9: Ausschnitt aus dem Index für Satzanfänge

Satzanfänge und -enden

Im Rahmen dieser Arbeit wurde festgestellt, dass sich die erstellte Tabelle für den Satzindex auch eignet, um Sätze mit vorgegebenen Satzanfang abzurufen. Das Satzpräfix wird dazu genauso wie die Anfragesätze beim Satzindex in eine Zeichenkette umgewandelt. In der Tabelle zum Satzindex wird dann eine Präfixsuche ausgeführt. Diese wird von der Datenbank effizient durchgeführt. Die Treffermenge wird durch die Vorgabe einer minimalen Tokenanzahl und der Sprache weiter reduziert. Analog zur Satzsuche müssen bei zu langer Anfrage die Ergebnisse, die der Index liefert, noch überprüft werden.

Für die Suche nach Satzenden wurde eine zweite Tabelle aufgebaut, die darin von der Satzindex-Tabelle unterscheidet, dass die Reihenfolge der Token vor der Erzeugung einer maximal 56 Zeichen langen Zeichenkette umgekehrt wird.

Grundformen und POS-Tags

Mit dem Modul für die Satzindizes können nicht nur Token indiziert werden. Auch die annotierten Grundformen und POS-Tags eignen sich. Abbildung 3.10 zeigt einen Ausschnitt aus dem Index für die Grundformfolgen am Satzende. Mit ihm können Sätze abgefragt werden, die auf eine vorgegebene Abfolge von Grundformen enden.

Bei den Grundformen tritt das Problem auf, dass je Token mehr als eine Grundform annotiert sein kann. Damit ein Satz mit jeder in Frage kommenden Grundformenfolge gefunden werden kann, muss jede mögliche Kombination in den Index aufgenommen werden. Die Anzahl der Kombinationen ist das Produkt der Anzahlen der Grundformen, die für jedes einzelne Token annotiert sind. Zwar weisen von den 271 907 deutschsprachigen Segmenten nur 1047 mehr als 16 Kombinationen auf. Aber einige Segmente weisen zwischen 12 288 und 134 217 728 Kombinationen auf. Betroffen sind vor allem große Segmente aus $n:1$ Alignment-Beads und Segmente, die umfangreiches Tabellenmaterial enthalten.

```
mysql> SELECT name, beschr1 AS '#', datum AS 'SegNr.'
    -> FROM lemmata_suffix_strict WHERE name LIKE '.|reputation%'
    -> AND beschr2 = 2;
+-----------------------------------------------------+----+--------+
| name                                                | #  | SegNr. |
+-----------------------------------------------------+----+--------+
| .|reputation|#s|agency|the|hurt|be|headline|result|and|s | 54 | 490844 |
| .|reputation|#s|master|her|to|and|,|master|her|to|due|be | 24 | 682710 |
| .|reputation|commercial|its|and|relation|customer|its|,| | 42 | 638803 |
| .|reputation|his|damage|to|campaign|a|of|victim|himself| | 13 | 486772 |
| .|reputation|horrible|a|get|be|it|and|#|,|<unknown>|<unk | 17 | 683204 |
| .|reputation|international|good|a|with|minister|finance| | 12 | 485322 |
| .|reputation|of|loss|a|and|donation|reduced|against|warn | 27 | 526437 |
| .|reputation|scientific|excellent|its|note|would|<unknow | 19 | 439104 |
+-----------------------------------------------------+----+--------+
8 rows in set (0.00 sec)
```

Abbildung 3.10: Ausschnitt aus dem Index für Grundformfolgen am Satzende

Um die Indizes für Grundformenfolgen an Satzanfängen und -enden in vertretbarer Zeit aufbauen zu können, werden nur soviele Grundformenlisten aufgeteilt, dass eine voreingestellte Maximalanzahl von Kombinationen (erst 192, später auf 32 reduziert) nicht überschritten wird. Eine Verbesserungsmöglichkeit wäre, jeweils zu prüfen, ob sich die Grundformalternativen überhaupt in den 56 tatsächlich indizierten Zeichen niederschlagen.

Teilmengen der Token eines Segments

Zum Finden von Fuzzy-Matches kann ein Satzindex nicht verwendet werden. Selbst wenn sowohl der Satzanfang- als auch der Satzendenindex verwendet wird, können Sätze nicht gefunden werdem, die am Anfang und Ende Unterschiede zum Anfragesatz aufweisen. Gewünscht ist, dass alle Sätze gefunden werden, die eine vorgegebene Anzahl von Token (oder Grundformen) mit dem Anfragesatz gemeinsam haben. Dieses Suchproblem ist bereits aus dem Information-Retrieval bekannt. In einem Translation Memory werden statt Dokumenten Sätze gesucht.

Mit den datenbankseitig vorhandenen Indizes kann die Suche nach Sätzen, die k Token von n gegebenen Token $T_1, \ldots, T_n$ enthalten, bereits durchgeführt werden, ohne die Sätze selbst aus der Datenbank auslesen zu müssen. Dazu werden für jede k elementige Teilmenge $T_{i_1}, \ldots T_{i_k}$ der Anfragetoken die Menge der Satznummern der Sätze ermittelt, die die jeweiligen k Token enthalten. Die Vereinigung dieser $\binom{n}{k}$ Mengen gibt die gesuchten Sätze an. Diese einzelnen Mengenoperationen gibt folgender Ausdruck wieder:

$$\bigcup_{1 \leq i_1 < \ldots < i_k \leq n} \bigcap_{j=1}^{k} R(T_{i_j}),$$

wobei R ein Token auf die Menge der Satznummern der Sätze abbildet, in denen das Token vorkommt. R kann mit einer einfachen SQL-Anfrage implementiert werden. Die Mengenoperationen können prinzipiell auch von der Datenbank ausgeführt werden. Im Rahmen

genoperationen können prinzipiell auch von der Datenbank ausgeführt werden. Im Rahmen dieser Arbeit[34] wurde jedoch darauf verzichtet, da der Autor keine Erfahrungen darin hat, ob die verwendete MySQL-Datenbank erkennt, dass hier viele Zwischenergebnisse wiederverwendetet werden können. Die Mengenoperationen werden anwendungsseitig im Fuzzy-Matching Modul ausgeführt.

Das Laufzeitverhalten ist sehr schlecht, wenn die Mengenoperationen wie oben notiert ausgeführt werden, da dann $\binom{n}{k}$ Schnittmengen bestimmt werden müssen. Liegen die Mengen $R(T_i)$ als sortierte Listen vor, dann kann in $O(n^2m)$ (m sei die Länge der längsten Liste, d. h. $m = \max|R(T_i)|$) bestimmt werden, welche Satznummern mindestens k mal auftreten. Dies wurde aber nicht implementiert, da eine Beschränkung von k auf $k \leq 3$ vertretbar erschien.

Anpassungen sind notwendig, wenn in der Anfrage Token mehrfach auftreten dürfen. Man kann weiterhin mit obigen Mengenoperationen arbeiten, wenn statt mit Token mit Paaren bestehend aus Token und Nummer des Auftretens im Satz gearbeitet wird. Ein entsprechender Index müsste dazu aufgebaut werden.

Ein anderer Ansatz wurde in der Zeit zwischen KoKS-Projekt und der Erstellung dieser Arbeit verfolgt. Es wurden alle zwei- und dreielementigen Teilmengen von Token indiziert, die in Sätzen des Korpus vorkommen. Motivation ist, dass die Mengen $R(T_i)$ sehr groß sein können. Mit dem zusätzlichen Index können Mengen $R(T_i) \cap R(T_j)$ und $R(T_i) \cap R(T_j) \cap R(T_o)$ direkt abgerufen werden.[35] Der Zeitbedarf für den Indexaufbau stellte sich jedoch als Problem heraus. Im Nachhinein kann vermutet werden, dass dies an den sehr langen Segmenten liegt, die beim Ausmultiplizieren der Grundformen bereits Probleme bereiteten. Alle beschriebenen Indizes wurden auch für die Suche mit Grundformen implementiert.

Anpassung für Grundformen und POS-Tags

Mit Grundformen oder POS-Tags kann auf gleiche Weise gesucht werden. Die notwendige Anpassung der Retrieval-Funktion R erfordert nur einen Rückgriff auf andere Tabellen. Zur Erinnerung: Die Token sind nicht direkt mit der Korpustabelle verknüpft, sondern stehen in einer Tokentupel-Tabelle bestehend aus Token, Grundform, POS-Tag und Sprache. Wenn die Zeichenketten der Token, Grundformen und POS-Tags auf genau gleiche Weise mit der Tokentupel-Tabelle verknüpft wären, müsste nur der Name einer Tabelle in den Datenbankanfragen ersetzt werden. Leider ist dies nicht der Fall. Die Token stehen direkt in der Tokentupel-Tabelle, die Grundformen in einer Extratabelle und die POS-Tags in mehreren Tabellen (je Tagset eine Tabelle).

Suche nach POS-Tagfolgen

Die Suche nach POS-Tagfolgen wurde vorbereitet, da erwartet wurde, dass sie für diese Arbeit interessant werden könnte. Soweit ist es aber nicht gekommen, sodass sie nicht implementiert wurde.

Ein spezieller Index ist sinnvoll, da ein einfacher Ansatz, der das Retrieval aus dem vorangehenden Unterabschnitt nutzt und dann die Ergebnisse danach filtert, ob die POS-Tags in der richtigen Reihenfolge und zusammenhängend auftreten, zwei Probleme aufwirft. Zum einen sind die Zwischenergebisse sehr umfangreich. Beispielsweise dürfte $R_{POS}('\mathrm{NN}')$ fast alle Satznummern des Korpus enthalten. Zum anderen dürfte auch das

[34] Im KoKS-Projekt wurde nur der Sonderfall $k = n$ implementiert, bei dem die Vereinigung entfällt.

[35] Durch eine geschickte Verteilung der k Anfragetoken auf $\lceil \frac{k}{3} \rceil$ Indexanfragen, die die Häufigkeit der Token gemessen am Gesamtkorpus berücksichtigt, kann man sehr kleine Ergebnismengen erhalten.

Korpus	Deutsch	Englisch	Verhältnis	Ausgangssprache
DE-News	7 045 756	6 502 884	1,08	Deustch
EU	24 167 152	21 050 021	1,15	unbekannt
Harry Potter	3 055 845	2 675 042	1,14	Englisch
Gesamt	34 268 753	30 227 947	1,13	—

Tabelle 3.4: Anzahl der Zeichen in den verwendeten Korpora

Endergebnis des Retrievals viele Sätze enthalten, die beim anschließenden Filtern verworfen werden müssen.

Aus dem Information-Retrieval ist der Ansatz bekannt, dass im Index zusätzlich zur Satznummer auch die Position des indizierten POS-Tags im Satz vermerkt wird. Die Reihenfolge und Kontinuität der POS-Tags kann dann ohne Auslesen der gesamten Sätze geprüft werden. Die Zahl der Überprüfung ändert sich damit aber nicht.

Wenn nicht einzelne POS-Tags, sondern alle Folgen von POS-Tags indiziert würden, könnte direkt im Index nachgeschlagen werden. Dies ist aber nicht praktikabel, da die Zahl der Sequenzen in einem Satz quadratisch von der Satzlänge abhängt. Mit einer Beschränkung auf kurze POS-Tagfolgen im Index kann dieses Problem gelöst werden. Die Anfrage kann weiterhin aus langen POS-Tagfolgen bestehen, wenn weiterhin nachgefiltert wird. Dazu muss die Anfragefolge in indexgerechte Stücke zerteilt werden. Freiheiten bei der Zerlegung könnten genutzt werden, um möglichst seltene POS-Tagfolgen für die Indexanfrage zu nutzen.

3.3 Eigenschaften

Die Größenangaben im KoKS-Abschlussbericht von Erpenbeck et al. (2002) beziehen sich auf das gesamte Korpus, das aufbereitet wurde. Bereits im KoKS-Projekt wurde nicht das ganze Korpus weiterverarbeitet. Nach Verbesserungen an einigen Komponenten wurde die Vorverarbeitung nochmal durchgeführt unter Verwendung von Zwischenergebnissen aus vorangegangenen Durchläufen. Dabei standen nicht für alle Teilkorpora die notwendigen Daten zur Verfügung, da nicht von Anfang an die Zwischenergebnisse gespeichert wurden und womöglich auch gespeicherte Ergebnisse gelöscht wurden, um Platz für neue Ergebnisse zu schaffen.[36] Einige Jahrgänge des EU-Korpus und der DE-News Nachrichten stehen daher nicht zur Verfügung, sodass eine Neuauszählung dieser Teilkorpora angebracht ist. Schließlich ist das Harry-Potter-Korpus neu hinzugekommen, für das im KoKS-Abschlussbericht keine Daten vorhanden sind.

3.3.1 Größe

Die Anzahl der Sätze wurde bereits in Tabelle 2.1 auf Seite 15 angegeben. Die Segmentanzahlen ergeben sich aus den Zeilensummen in der Tabelle 2.2. Die Summen sind 57 599, 101 828 und 33 377 für die Teilkorpora „DE-News“, „EU“ und „Harry Potter“.

[36] Es sind keine Protokolle vorhanden, aus denen der genaue Ablauf der Vorverarbeitung für jeden Teilkorpus rekonstruiert werden könnte.

Korpus	Deutsch	Englisch	Verhältnis	Ausgangssprache
DE-News	884 130	1 026 389	0,86	Deustch
EU	2 992 002	3 166 040	0,95	unbekannt
Harry Potter	475 189	464 690	1,02	Englisch
Gesamt	4 351 321	4 657 119	0,93	—

Tabelle 3.5: Anzahl der Wörter in den verwendeten Korpora

Korpus	Deutsch	Englisch	Verhältnis	Ausgangssprache
DE-News	961 104	1 119 728	0,86	Deustch
EU	3 309 335	3 493 419	0,95	unbekannt
Harry Potter	588 905	584 117	1,01	Englisch
Gesamt	4 859 344	5 197 264	0,93	—

Tabelle 3.6: Anzahl der Token in den verwendeten Korpora

Tabellen 3.4 bis 3.6 zeigen analog die Anzahl der Zeichen, Wörter[37] und Token. Zeichen und Wörter wurden in den aufbereiten, aber noch nicht tokenisierten Dateien gezählt. Die Zahlen können nicht mit der in der Datenbank vorliegenden Token- und Satzanzahl gleichgesetzt werden, da manche Dokumente nicht vom Aligner verarbeitet werden konnten.[38] Beim EU-Korpus sind etwa 1,5 % der Dateien betroffen, bei den DE-News nur 0,3 %. Das Harry-Potter-Korpus konnte vollständig verarbeitet werden.

3.3.2 Frequente Wörter

Die Häufigkeiten, mit denen Wörter im Korpus auftreten, geben einen Anhaltspunkt, welche Themen oder Themenfelder dominieren. Eine kurze Liste der häufigsten Wörter reicht hierzu aber nicht aus. In den höchsten Rängen stehen fast ausschließlich Artikel, Präpositionen und Satzzeichen. Beispielsweise steht das Token „der“, das 174 292 mal im Korpus auftritt, auf Rang drei hinter den Satzzeichen Komma und Punkt.

Inhaltstragend sind Wörter offener Wortklassen. Tabelle 3.7 zeigt die häufigsten Token, die als gewöhnliches Nomen[39] getaggt wurden. Die Liste bestätigt, dass das Korpus hauptsächlich aus EU-Dokumenten besteht, die den politischen Rahmen der wirtschaftlichen Zusammenarbeit beschreiben. In der Rangliste der Eigennamen, die hier nicht abgebildet ist, steht der Name „Harry“ auf dem ersten Rang vor „ECU“. Dies zeigt, dass auch kleine Teilkorpora einen Einfluss auf das Gesamtkorpus haben können, wenn sie ungewöhnliche Merkmale aufweisen.

Hier wurden die Häufigkeiten der Token ermittelt. Flektierte Formen und Großschreibungen am Satzanfang werden dadurch als eigenes Wort aufgeführt. Will man die verschiedenen Formen eines Wortes zusammenfassen, dann müssen statt der Token die zugehörigen

[37] Shell-Kommando `wc -wc`

[38] Die eingangs genannten Segmentanzahlen können sich nur auf die vollständig verarbeiteten Dokumente beziehen, da Segmente erst im letzten Verarbeitungsschritt, dem Alignment, gebildet werden.i

[39] IMS Tagset und Penn-Treebank Tagset unterscheiden zwischen Eigennamen und allen anderen, „normalen“ Nomen. Das Penn-Treebank Tagset enthält zusätzlich Nomen-Tags mit dem Suffix „S“, die verwendet werden, um im Plural stehende Nomen zu kennzeichnen.

Deutsch		Englisch	
Häufigkeit	Token	Häufigkeit	Token
25485	Kommission	10981	%
10410	%	9784	aid
7950	Gemeinschaft	8404	market
7184	Unternehmen	6925	something
6558	Mitgliedstaaten	6430	programme
5904	Maßnahmen	5959	development
5721	Entwicklung	5498	time
4977	Rahmen	5425	countries
4562	Jahr	5421	year
3383	Programm	5119	measures
3293	Zusammenarbeit	4900	somebody
3120	Hilfe	4547	policy

Tabelle 3.7: Häufige Token mit POS-Tags ’NN’ und ’NNS’

Grundformen ausgezählt werden. Hierbei können aber Wörter nicht berücksichtigt werden, die der Lemmatisierer nicht kennt. Auf eine Darstellung dieser Häufigkeiten wird hier verzichtet, da hier nur ein grober Eindruck zur Unausgewogenheit des Korpus vermittelt werden soll. Dazu sollte Tabelle 3.7 reichen.[40]

3.3.3 Alignment

Daten zu dem Alignment wurden bereits im Abschnitt 2.2.3 vorgestellt. Siehe insbesondere Tabelle 2.2 auf Seite 16.

3.4 Belegsituation

In diesem Abschnitt soll beleuchtet werden, wie gut die Aussichten sind, in dem verwendeten Korpus Material zu finden, das bei der Übersetzung eines neuen Satzes hilft.

3.4.1 Stichprobe

Je Sprache (Deutsch und Englisch) wurden mindestens 250 Segmente ausgewählt. Es wurde darauf verzichtet, die Auswahl durch einen (Pseudo-) Zufallsprozess zu steuern. Stattdessen wurden Segmente ausgewählt, deren Segment-Nummer sich ohne Rest durch eine zuvor bestimmte Zahl teilen läßt. Da zusätzlich die Länge der Sätze auf 12 bis 60 Wörter eingeschränkt wurde und da die fortlaufende Nummerierung der Korpussegmente zwischen den Dokumenten durch die Segmente der parallelen Sprache unterbrochen wird, kann die Zahl der ausgewählten Segmente nur ungenau mit dem Teiler gesteuert werden. Solange weniger als die gewünschten 250 Segmente in der Stichprobe enthalten sind, wird für die

[40] Die jeweils tausend häufigsten Token, Lemmata und POS-Tags stehen im Quellcode des Moduls `DatabaseAPI/haeufigkeit.py`.

verbleibende Anzahl ein neuer Teiler bestimmt und der Auswahlprozess wiederholt. Eine zu große Auswahl wurde nicht reduziert, da dies unnötig erschien.[41] Auf diese Weise wurden 250 Segmente im Deutschen Korpusteil und 260 Segmente im Englischen Korpusteil als Stichprobe bestimmt.

In Folgendem wird wie schon in anderen Abschnitten vereinfachend von Sätzen der Stichprobe gesprochen, obwohl Segmente mehr als einen Satz enthalten können.

3.4.2 Ermittlung der Fuzzy-Matches

Zu jedem der 510 Sätze der Stichprobe werden 11 Fuzzy-Matches aus der Datenbank abgefragt. Da der Anfragesatz selbst in der Datenbank vorhanden ist, sind unter den Treffern zehn neue Fundstellen.[42]

In Folgendem wird beschrieben, wie die Fuzzy-Matches ermittelt werden. Wie bereits in Abschnitt 2.3.1 erwähnt, habe ich keine Literatur zu diesem Spezialthema gesucht. Eine Implementation des im Abschnitt 2.3.1 skizzierten Ansatzes schien mit den im KoKS-System vorhandenen Komponenten leicht umsetzbar zu sein. Wie in der nachfolgenden Darstellung deutlich wird, mussten jedoch mehrere Detailprobleme gelöst werden.

Einschränkung der Kandidaten

Im ersten Teil der Fuzzy-Match-Suche wird die Kandidatenmenge soweit eingeschränkt, dass nur ein kleiner Teil des Gesamtkorpus genauer geprüft werden muss. Im wesentlichen wird dazu die im Abschnitt 3.2.7 Methode zum Zugriff auf Sätze, die eine Teilmenge der Token des Anfragesatzes enthalten, verwendet.

Expansion der Anfragetoken Verwendet man nur die Token des Anfragesatzes für die Suche im Korpus, dann werden Abweichungen in der Flexion genauso behandelt wie Ersetzungen durch andere Wörter. Hat der Satz sonst nicht genug Wörter mit dem Anfragsatz gemeinsam, wird er nicht in die Menge der Kandidat aufgenommen. Ein solcher Fall kann beispielsweise eintreten, wenn das Subjekt eines kurzen Satzes den Numerus wechselt. Verb, Nomen, Artikel und Adjektive, die zum Subjekt gehören, können sich dann geringfügig verändern, sodass die Sätze auf Tokenebene wenig oder nichts gemeinsam haben.

Das Problem könnte leicht mit einer Suche mittels der annotierten Grundformen gelöst werden, wenn die Annotation eindeutig und vollständig wäre. Der Tagger annotiert jedoch Grundformlisten, wenn die Grundform nicht eindeutig aus dem Lexikon des Taggers hervorgeht, oder gar keine Grundform, wenn die Vollform unbekannt ist. (Siehe auch Abschnitt 3.2.3.) Im letzteren Fall kann nur mit dem Token gesucht werden. Der erste Fall kann sowohl im Anfragesatz als auch im Korpus auftreten. Für die Suche werden daher sämtliche Grundformlisten zusammengestellt, die eine Grundform enthalten, die in der Liste der Grundformen des Anfragetokens vorkommen. (Eindeutige Grundformannotationen werden dabei als einelementige Listen behandelt.) Beispielsweise werden zum Token „fiel“ die drei Grundformlisten „fallen“, „fallen, gefallen“ und „fallen, fällen“ gebildet. Diese Aufgabe wird mit der im Abschnitt 3.2.6 beschriebenen Grundformentabelle effizient durchgeführt.

Da bei der Abfrage von Fundstellen zu Grundformlisten grundsätzlich die Tokentupeltabelle verwendet wird, übersetzt die Datenbank implizit jede Grundformliste in die Menge

[41] Dies wäre jedoch leicht zu realisieren gewesen und hätte die Auswertung vereinfacht.

[42] Der Anfragesatz muss nicht unter den ersten elf Treffern sein, wenn mindestens zwölf Exact-Matches vorhanden sind. Bei der Stichprobe trat dieser Fall aber nicht auf.

der Token, die mit ihr annotiert wurden. Es wird also für jedes Anfragetoken mit einer Menge von Token nach Fundstellen gesucht. Im Fall, dass direkt mit dem Anfragetoken gesucht wird, ist die Menge einelementig. Die Menge enthält aber auch im anderen Fall immer das Anfragetoken.[43] Daher wird dieser Schritt hier als Expansion der Anfragetoken bezeichnet.

Ermittlung der Häufigkeiten Als nächstes wird für jedes expandierte Anfragetoken die ungefähre Häufigkeit im Korpus ermittelt, um zu entscheiden, welche Token für die Suche im Korpus benutzt werden. Die Häufigkeit des Anfragetokens und der zusammengestellten Grundformen wird in Tabellen der häufigsten tausend Token bzw. Grundformen nachgeschlagen. (Die Werte sind nicht exakt, da die Tabellen nicht auf dem aktuellen Stand des Korpus sind.) Ist keine der Formen in den Häufigkeitstabellen gelistet, wird die Häufigkeit null unterstellt. Sie wird hier verwendet, um seltene Token zu kennzeichnen, und bedeutet nicht etwa, dass das Token nicht im Korpus aufträte.

Auswahl der Anfragetoken Ein expandiertes Token wird für die Suche im Korpus herangezogen, wenn dessen Häufigkeit unter einem Schwellwert liegt, der in etwa die häufigsten 200 Token ausschließt. Wenn weniger als acht Token ausgewählt werden, wird die Schwelle abhängig von der bisherigen Anzahl der selektierten Token moderat erhöht. Nur wenn die Anzahl trotzdem unter zwei bleibt, wird die Schwelle so weit erhöht, dass selbst Formen von „sein“, „werden“ (Deutsch) und „have“ (Englisch) ausgewählt werden.

Die Beschränkung auf nicht zu häufige Token hat große Ähnlichkeit mit der Verwendung von so genannten Stoppwortlisten, die nicht zu berücksichtigende Wörter benennen. Hier würde eine solche Liste alle Wörter enthalten, die keinen Beitrag zur Einschränkung der Kandidatenmenge erwarten lassen. Der Unterschied des hier gewählten Auswahlverfahrens zu Stoppwortlisten ist die Anpassung der Häufigkeitsschwelle an die Zahl der bisher aufgenommenen Token. Beispielsweise werden zu der Anfrage „Sein oder nicht sein.“ die Anfragetoken „oder“ und „nicht“ verwendet, obwohl sie auf den Häufigkeitsrängen 109 und 47 stehen.[44] Der beste Fuzzy-Match „Sein oder Nichtsein“ wird in den Wörterbüchern des KoKS Systems gefunden. (Zur Berwertung der Güte eines Treffers siehe weiter unten.) Der zweitbeste Treffer „Oder nicht?“ stammt aus dem Harry Potter Korpus (Band 4, Kapitel „Der Todesser“). Mit einer Stoppwortliste hätte kein expandiertes Token für den Korpuszugriff zur Verfügung gestanden, sodass die Treffermenge leer gewesen wäre.

Wahl der Mindestanzahl der Übereinstimmungen Die Zahl k, die angibt, wie viele der n ausgewählten Anfragetoken in einem Satz vorkommen müssen, damit er in die Kandidatenmenge für die Fuzzy-Matches aufgenommen wird, ist der zweite Faktor, der die

[43] Genau genommen müsste man hier von den Tokentupel-IDs sprechen. Unter der Annahme, dass der IMS TreeTagger ein Token, das er einmal lemmatisieren konnte, nie mit „`<unknown>`“ annotiert, deckt die erstellte Grundformliste alle Tokentupel ab, in denen das Token auftritt. Im Bezug auf das Anfragetoken ist die Darstellung also korrekt. Jedoch ist die Vorstellung falsch, die Token, die mit einer Grundformliste aus der Liste der Grundformenlisten annotiert wurden, würden für die Suche im Korpus verwendet. Im Beispiel zu „fiel“ wird dies deutlich: Obwohl im Korpus das Token „Gefallen“ achtmal mit der Grundformliste „fallen, gefallen“ annotiert wurde, werden die anderen 29 Auftreten von „Gefallen“ bei einer Suche mit der Grundformliste ignoriert, da hier das Nomen vorliegt.

[44] Diese Rangzahlen müssen in etwa halbiert werden, wenn sie mit einsprachigen Häufigkeitstabellen verglichen werden, da in KoKS die Häufigkeiten sprachübergreifend ausgezählt wurden. Die Auszählung und Verwendung der Tabelle ist auf diese Weise einfacher. Ein Problem sei aber nicht verschwiegen: Bei Token, die in beiden Sprachen auftreten, ist die so bestimmte Häufigkeit die Summe der Häufigkeiten in den Einzelsprachen. Unter den häufigsten 200 Token sind hier besonders Satzzeichen und die drei Token „Union“, „national“ und „international“ betroffen. Sie stehen auf zu hohen Rangplätzen. (Token wie „Land“ sind nur geringfügig betroffen, da sie im Englischen selten groß geschrieben werden.)

Auswahl der Kandidaten steuert. Je kleiner k gewählt wird, desto mehr Sätze werden als Fuzzy-Match in Betracht gezogen.

Für ein Translation Memory, das nur ganze Sätze mit geringem Korrekturbedarf als Übersetzungsvorschläge anbieten will, würde es Sinn machen, nur eine feste Anzahl von Abweichungen zu erlauben. Wenn beispielsweise maximal zwei Wörter unterschiedlich sein dürfen, könnte man $k = n - 2$ wählen. Dagegen muss ein kleiner Wert für k eingesetzt werden, wenn auch Sätze mit wenigen Übereinstimmungen gefunden werden sollen. Ein sehr kleiner Wert, z. B. $k = 3$, könnte zum Auffinden von kurzen Satzfragmenten, so genannte Subsegment-Matches, dienen. Hierbei ist wichtig, dass häufige Token zuvor von der Suche ausgeschlossen wurden, da sonst viele irrelevante Sätze gefunden werden, die nur in Artikeln, Präpositionen, Konjunktionen oder anderen häufigen Wörtern mit dem Anfragesatz übereinstimmen.[45]

Für die Fuzzy-Matches der Stichprobe wurde $k = \min(3, \lceil \frac{n}{2} \rceil)$ gesetzt, um die Anzahl der auszuführenden Korpusanfragen klein zu halten. Sie liegt in $O(n^3)$, da n in der Anzahl $\binom{n}{k}$ mit $k \leq 3$ höchstens in der dritten Potenz auftreten kann. Mit dieser Wahl von k werden viele Sätze als Kandidaten zugelassen. Nur selten sollten also relevante Sätze nicht enthalten sein. Die zusätzliche Zeit, die die Verarbeitung der großen Kandidatenmenge erfordert, ist hier anders als in einer interaktiven TM-Anwendung kein Hindernis.

Korpuszugriff Zu jedem ausgewählten Anfragetoken werden zuerst die Satznummern der Sätze bestimmt, in denen eine Form des expandierten Tokens auftritt. Jeweils k Satznummerlisten werden dann geschnitten, um die Sätze zu ermitteln, in denen mindestens k Anfragetoken vorkommen. Die Vereinigung aller $\binom{n}{k}$ Schnitte ergibt schließlich die Kandidatenmenge. Auf Seite 52 im Abschnitt 3.2.7 sind diese Operationen als Formel notiert. Im Abschnitt 2.3.1 ist der Spezialfall mit $k = 1$ beschrieben.

Bewertung mit Ähnlichkeitsmaß

Aus der Kandidatenmenge können die Sätze, die als Fuzzy-Matches gelten sollen, mit aufwendigeren Methoden ausgewählt werden, da diese Menge wesentlich kleiner ist als das Gesamtkorpus. Im Abschnitt 2.3.2 werden Möglichkeiten angedeutet, wie linguistisches Wissen in die Bewertung der Relevanz der Kandidaten einbezogen werden kann, und auf Baldwin und Tanaka (2000) verwiesen, die mehrere Ähnlichkeitsmaße daraufhin untersuchen, wie sie die Qualität der Übersetzungsvorschläge in einer TM-Anwendung beeinflussen.

Auch hier wird ein Ähnlichkeitsmaß, das den Grad der Übereinstimmung von Anfragesatz und Kandidat bestimmt, als Maß der Relevanz verwendet. Es handelt sich um ein einfaches, zeichenbasiertes Maß, das im KoKS-Projekt entwickelt wurde. Das Ähnlichkeitsmaß stützt sich nicht auf einzelne Zeichen, sondern auf alle Sequenzen von drei Zeichen, die im Satz auftreten. Diese Sequenzen nennt man Trigramme. Seien $c_1(t)$ und $c_2(t)$ die Häufigkeiten der Trigramme t in den zu vergleichenden Zeichenfolgen. Dann wird als Ähnlichkeit der Wert

$$a = \frac{\sum_t \min(c_1(t), c_2(t))}{\sum_t \max(c_1(t), c_2(t))}$$

[45] Alternativ könnte man nach der Bildung der k elementigen Teilmengen der Anfragetoken diejenigen ausfiltern, die zu wenig seltene Wörter enthalten, um eine kleine Kandidatenmenge erwarten zu können. Zusätzlich könnte man verlangen, dass die Token im Anfragesatz eng zusammenstehen. (Die gleiche Bedingung könnte man auch an die zu findenen Sätze knüpfen. Mit den vorhandenen Indizes kann dies aber nicht effizient durchgeführt werden.) So wäre es möglich, nach Sequenzen von Wörtern offener und geschlossener Wortklassen, wie z. B. „im Schatten der Bäume“, zu suchen.

	1	2	3	4	5	6	7
1	100 %	20 %	10 %	24 %	3 %	4 %	3 %
2	20 %	100 %	0 %	37 %	21 %	27 %	21 %
3	10 %	0 %	100 %	3 %	18 %	24 %	18 %
4	24 %	37 %	3 %	100 %	29 %	35 %	29 %
5	3 %	21 %	18 %	29 %	100 %	58 %	44 %
6	4 %	27 %	24 %	35 %	58 %	100 %	69 %
7	3 %	21 %	18 %	29 %	44 %	69 %	100 %

1: Baumes, 2: Baumschatten, 3: Bäume, 4: der Schatten eines Baumes, 5: der lange Schatten der Bäume, 6: im Schatten der Bäume, 7: im Schatten der großen Bäume

Tabelle 3.8: Ähnlichkeitswerte für einige kurze Zeichenfolgen

eingesetzt.[46] Der Wert liegt zwischen null und eins. Für identische Zeichenfolgen ist die Ähnlichkeit eins, d. h. 100 %.

Zu den nachfolgend angegebenen Zahlen sollte erwähnt werden, dass Leerzeichen am Anfang und Ende der Zeichfolgen hinzugefügt und die Klein-/Großschreibung und Satzzeichen ignoriert werden. Im KoKS-Abschlussbericht auf Seite 57 bis 62 beschreiben Erpenbeck et al. (2002) das Ähnlichkeitsmaß detailiert und geben viele Beispiele für den bilingualen Anwendungsfall an.[47] Hier sind neue Beispiele nötig, da die zu vergleichenden Sätze bei der Fuzzy-Match Suche einsprachig sind. Tabelle 3.8 zeigt für sieben Zeichenfolgen die Ähnlichkeitswerte aller Paare. Beispielsweise wird den Zeichenfolgen „Baumes“ und „Baumschatten“ eine Ähnlichkeit von $\frac{3}{15} = 20\%$ zugeschrieben. (Drei von 15 Trigrammen, nämlich „ ba“, „bau“ und „aum“, treten in beiden Zeichenfolgen auf.) Die Tabelle ist symmetrisch um die Diagonale, da das Maß symmetrisch ist.

Beispiel

Der konstruierte Satz

(3.1) Im langen Schatten eines großen Baumes kann man sehr gut Spinnen fangen.

soll hier als Beispiel dienen. Er wurde so gewählt, dass unterschiedliche Kombinationen von Anfragetoken zu Fuzzy-Match-Kandidaten führen. Folgende acht Token werden ausgewählt: 0: Schatten, 1: Baumes, 2: Spinnen, 3: fangen, 4: langen, 5: sehr, 6: man, 7: gut. Angegeben sind Indexnummern, die im Folgenden verwendet werden. Die Reihenfolge der Token ist aufsteigend mit der festgestellten Häufigkeit. Es wird $k = \min(3, 8/2) = 3$ gesetzt. Nur neun der $\binom{8}{3} = 56$ möglichen Kombinationen von Anfragetoken ergeben Kandidaten:

[46] Baldwin und Tanaka (2000) geben ein zeichenbasiertes Maß „Token Intersection“ an, dass im Nenner anstatt des Maximums das arithmetische Mittel verwendet (Formel 2 auf Seite 38). Im Nenner steht der Mittelwert der Längen der Zeichenfolgen. Der Bruch wurde mit zwei erweitert. Das sei die übliche Form. Wenn man für die Längen $\sum c_i(t)$ einsetzt und die Summen zusammenfasst, wird die Ähnlichkeit zum KoKS-Maß offensichtlich. Baldwin und Tanaka erlauben zusätzlich, dass der Einfluss jedes Tokens unterschiedlich gewichtet wird.

[47] Dort wird das Abstandsmaß $1 - a$ betrachtet.

```
Schnitt [0, 1, 2]: 2 Kandidaten
Schnitt [0, 1, 4]: 2 Kandidaten
Schnitt [0, 4, 6]: 1 Kandidat
Schnitt [1, 4, 7]: 1 Kandidat
Schnitt [2, 3, 4]: 2 Kandidaten
Schnitt [4, 5, 6]: 1 Kandidat
Schnitt [4, 5, 7]: 2 Kandidaten
Schnitt [4, 6, 7]: 1 Kandidat
Schnitt [5, 6, 7]: 4 Kandidaten
Vereinigung: 16 Kandidaten
```

Hier ist die Summe der Anzahlen in den einzelnen Schnitten gleich der Mächtigkeit der Vereinigung. Dies ist ungewöhnlich und bedeutet, dass die Schnitte paarweise disjunkt sind. Mit $k = 4$ hätte man zu dem Beispielsatz folglich keine Kandidaten erhalten.

Wie für die Stichprobe werden die elf Kandidaten mit der größten Ähnlichkeit zum Anfragesatz als Fuzzy-Matches übernommen. Tabelle 3.9 zeigt die Fuzzy-Matches geordnet nach Relevanz. Von den fünf übrigen nicht aufgeführten Kandidaten sind vier wesentlich länger als der schon lange elfte Fuzzy-Match und stammen aus dem EU Korpus. Der 16. Kandidat ist zwar kurz, hat aber nur „sehr“, „gut“ und „man“ mit dem Anfragesatz gemeinsam.

Die Fuzzy-Matches (und auch die übrigen Kandidaten) zu diesem Beispielsatz sind nur sehr eingeschränkt oder gar nicht für die Übersetzung des Anfragesatzes nützlich. Lediglich die Phrasen „Schatten der Bäume“ und „die langen Schatten der Bäume“ lassen irgendeine Hilfe erwarten, die über eine reine Einzelwortübersetzung hinausgeht. Da selbst solche Kandidaten noch von der entwickelten Fuzzy-Match-Suche ermittelt werden, kann man hoffen, dass nur sehr wenige relevante Korpusstellen übersehen werden.[48]

3.4.3 Klassifikation der Fuzzy-Matches

Die vielen Fuzzy-Matches — insgesamt sind es 4 476 — müssen genauer betrachtet werden, um die Frage nach der Belegsituation beantworten zu können. Es soll bestimmt werden, wie viele Fuzzy-Matches tatsächlich relevant sind. Interessant ist weiter, welche Arten von Fuzzy-Matches wie häufig auftreten. Dazu werden weiter unten Klassen definiert, die zur Annotation der Fuzzy-Matches verwendet werden sollen.

Alle Fuzzy-Matches manuell zu beurteilen wäre mit einem großen Zeitaufwand verbunden. Dies könnte vermieden werden, wenn sich herausstellt, dass das Ähnlichkeitsmaß die Relevanz gut vorhersagt.

Bearbeitungsreihenfolge

Um möglichst früh ein Bild über den Zusammenhang von Ähnlichkeit und Klasse der Fuzzy-Matches zu erhalten, werden zuerst solche mit möglichst verschiedenen Ähnlichkeitswerten klassifiziert. Dann werden Lücken in Wertebereichen geschlossen, in denen unterschiedliche Klassen auftreten.

Die absoluten Ähnlichkeitswerte sind ungeeignet für die Auswahl, da kleine Werte dominieren. Verwendet wird der Rang in der sortierten Liste der Werte. Das bedeutet, dass in

[48] Überprüfen könnte man dies, indem man k weiter absenkt. Mit $k = 2$ kommen im Beispiel 331 neue Kandidaten hinzu. Es tritt ein neuer Fuzzy-Match auf, der eine Ähnlichkeit von 23 % zum Anfragesatz hat und die Phrase „im Schatten eines Baumes“ enthält. Es werden also tatsächlich Stellen im Korpus übersehen. Man darf aber weiter hoffen, dass es nicht viele sind.

a	Fuzzy-Match	Quelle
18 %	Zwei einzelne Spinnen entflohen dem Licht des Zauberstabs in den Schatten der Bäume.	H.P. Bd 2
16 %	Und so folgten sie den huschenden Schatten der Spinnen in das Dickicht der Bäume.	H.P. Bd 2
14 %	Zwanzig Minuten lang gingen sie durch den Wald, laut redend und scherzend, bis sie endlich auf der anderen Seite zwischen den Bäumen hervortraten und sich im Schatten eines gigantischen Stadions fanden.	H.P Bd 4
14 %	Die letzten Strahlen der untergehenden Sonne tauchten das Land und die langen Schatten der Bäume in blutrotes Licht.	H.P. Bd 3
14 %	Man sieht, auch in der Wettbewerbspolitik wirft die WWU ihren langen und wohltünden Schatten voraus.	EU 1991
12 %	Dann verstecken wir uns am besten hinter einem Baum und halten Ausschau. # # Gut, aber hinter den Gewächshäusern lang!	H.P. Bd 3
12 %	Ron hatte ihm den ganzen Abend lang Ratschläge erteilt, zum Beispiel: # Wenn er versucht, dir einen Fluch anzuhängen, dann weich ihm besser aus, ich weiß nämlich nicht, wie man sie abblocken kann.	H.P. Bd 1
12 %	Moody langte in das Glas, fing eine Spinne ein und legte sie auf seinen Handballen, so daß alle sie sehen konnten.	H.P. Bd 4
10 %	Außerdem will sie auch ihre Bemühungen für eine bessere Ausbildung des Personals verstärken. Da in der Gemeinschaft bereits sehr lange Kernkraftwerke bestehen, wurden umfangreiche Betriebserfahrungen gesammelt, die ein beträchtliches Kapital darstellen.	EU 1990
8 %	# Der Kobold las den Brief sorgfältig durch. # Sehr gut #, sagte er und gab ihn Hagrid zurück. # Ich werde veranlassen, daß man Sie in beide Verliese führt.	H.P. Bd 1
7 %	Und dann fing er an, ihnen alles zu erzählen. Fast eine Viertelstunde lang sprach er in das gespannte Schweigen hinein: Er erzählte von der körperlosen Stimme und wie Her- mine schließlich begriffen hatte, daß er einen Basilisken in den Rohren gehört hatte; wie er und Ron den Spinnen in den Wald gefolgt waren, wo Aragog ihnen sagte, wo das letzte Opfer des Basilisken gestorben war; wie er auf den Gedanken kam, daß die Maulende Myrte dieses Opfer gewesen war und daß der Eingang zur Kammer des Schreckens in ihrer Toilette sein könnte...	H.P. Bd 2

Tabelle 3.9: Fuzzy-Matches zum Beispielsatz

Klasse	Relevanz
Exact-Match	100 %
nur Tippfehler	95 %
gleicher Inhalt	90 %
fast gleicher Inhalt	85 %
enthält etwas mehr	80 %
enthält etwas weniger	75 %
ähnlicher Inhalt	70 %
Subsegment-Match	60 %
Term-Match	20 %
keine Relevanz	0 %

Tabelle 3.10: Klassifikation der Fuzzy-Matches

der ersten Arbeitsphase zwischen je zwei zur Klassifikation ausgewählten Fuzzy-Matches in etwa gleich viele nicht ausgewählte liegen sollen.

Da der Vergleich von Fuzzy-Match und Anfragesatz das aufmerksame Lesen der Sätze erfordert, ist es sinnvoll, alle Fuzzy-Matches zu einem Satz der Stichprobe in einen Arbeitsgang zu annotieren. Prinzipiell erfordert dies keine Änderung am obigen Auswahlverfahren. Jedoch sind zu einem Satz der Stichprobe die jeweils besten Matches besonders interessant. Es wurde daher für jeden Satz der Stichprobe der Durchschnitt der Ähnlichkeitswerte der besten vier Fuzzy-Matches bestimmt und die Auswahl auf Grundlage dieser Werte vorgenommen.

Klassen

Wo die Grenze zwischen relevanten und irrelevanten Fuzzy-Matches zu ziehen ist, kann nicht im Allgemeinen beantwortet werden und hängt von der Anwendung ab. Im Falle eines Translation Memorys spielt die Arbeitsweise des Übersetzers sicherlich eine Rolle.

Verschiedene Klassen von Fuzzy-Matches können unterschieden werden. Bereits eingeführt wurden Exact-Match und Subsegment-Match. Für einen Subsegment-Match wird hier verlangt, dass eine Folge von mindestens acht Token übereinstimmt. Kürzere Subsegmente können einen Term-Match begründen, wenn es sich um einen gebräulichen Ausdruck, eine Kollokation oder einen Fachausdruck handelt. Ansonsten wird der Inhalt betrachtet.[49] Tabelle 3.10 listet die festgelegten Klassen auf. Die Grenzen zwischen den vier Klassen von „fast gleicher Inhalt“ bis „ähnlicher Inhalt“ sind schwer zu ziehen und haben sich während der Klassifikationsarbeit vermutlich verschoben. Die Klassen bilden keine lineare Skala. Beispielsweise sind die Pole „enthält mehr/weniger“ unabhängig vom Grad der inhaltlichen Ähnlichkeit.

Um die Klassifikation leichter mit den Trigramm-Ähnlichkeitswerten vergleichen zu können, werden den Klassen die in der Tabelle angegebenen Relevanzwerte zugewiesen. Als Indikator für die Güte der Belegsituation wird der Mittelwert der Relevanzwerte der besten vier Fuzzy-Matches jedes Stichprobensatzes bestimmt.

[49] Es wurde in Erwägung gezogen, die Klasse „ähnliches Subsegment“ nachträglich einzuführen. Da schon viele Fuzzy-Matches klassifiziert waren, wurde dies aufgegeben.

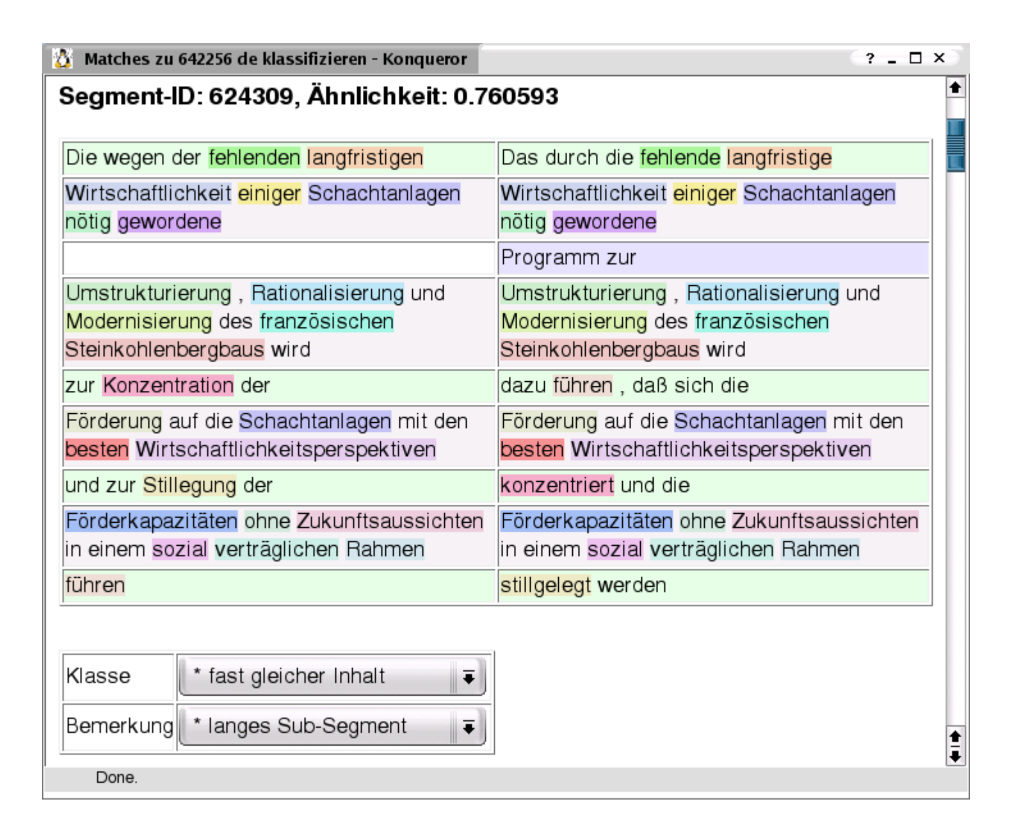


Abbildung 3.11: Annotationstool

Annotationstool

Für die Annotation der Klassen wurde eine webbasierte Anwendung implementiert, die ein sequentielles Wortalignment von Anfragesatz und Fuzzy-Match tabellarisch präsentiert und vom Benutzer die Klassifikation entgegennimmt. Zusätzlich werden mit dem Trigramm-Ähnlichkeitsmaß auffindbare Entsprechungen (einschließlich Überkreuzungen) farblich hervorgehoben. Der Annotator kann Übereinstimmungen an den Tabellenzeilen und Farbmustern schnell erkennen und kann sich so auf den Inhalt der Sätze konzentrieren.

Das Tool identifiziert automatisch Exact-Matches, Subsegment-Matches, Matches, die nur wenige Tippfehler enthalten, und Matches ohne Relevanz. Der Benutzer bestimmt aber grundsätzlich die Klasse eines Fuzzy-Matches im Dialog mit der Software. Das Tool unterbreitet nur Klassifikationsvorschläge. Abbildung 3.11 zeigt die Anwendung mit einem Fuzzy-Match aus dem EU-Teilkorpus. Die linke Tabellenspalte enthält den Anfragesatz. Rechts sieht man den zu klassifizierenden Fuzzy-Match. In den Zeilen der Tabelle werden abwechselnd unterschiedliche und übereinstimmende Tokenfolgen gegenübergestellt. Verschiedene Hintergrundfarben, die im Schwarzweißdruck gleich erscheinen, kennzeichnen die Zeilen. Die Hervorhebung einzelner Wörter können dagegen auch im Druck teilweise erkannt werden. In der farbigen Darstellung springen Entsprechungen wie von „Konzentration“ und „konzentriert“ und von „Stillegung“ und „stillgelegt“ schnell ins Auge, da nur

gefundene Matches	Häufigkeit	Deutsch	Englisch
0	17	15	2
1	16	14	2
2	10	6	4
3	11	5	6
4	4	2	2
5	10	7	3
6	12	10	2
7	6	4	2
8	4	2	2
9	5	3	2
10	415	182	233
Summe	510	250	260

Tabelle 3.11: Häufigkeiten der Anzahlen der Fuzzy-Matches

das Fehlen einer Farbe in einer Zeile entdeckt und die Stelle ihres Auftretens gefunden werden muss. Unten links wird in einer Auswahlliste die Klasse annotiert. Das Sternchen zeigt an, dass der Fuzzy-Match bereits annotiert wurde. Der Benutzer kann beliebig oft die Klassifikation ändern. Innerhalb der Fuzzy-Matches eines Stichprobensatzes kann mit dem Rollbalken (rechts außen) gewechselt werden. Für die einzelnen Sätze der Stichprobe gibt es eine Auswahlseite.

Ein Nachteile dieser Form der Annotation sollen nicht verschwiegen werden, nämlich die Reihenfolge, in der die Fuzzy-Matches zur Klassifikation vorgelegt werden. Innerhalb eines Anfragesatzes sind die Matches absteigend nach ihrer Ähnlichkeit zum Anfragesatz geordnet. Dies kann dazu verleiten, eine entsprechende Ordnung der Klassen anzunehmen. Besser wäre es, die Reihenfolge zu randomisieren. Mit diesem Problem einher geht die Anzeige der Ähnlichkeitswerte. Sie sollte während der Annotationsarbeit unterbleiben.

3.4.4 Ergebnisse

4 476 Fuzzy-Matches wurden zu den 510 Sätzen der Stichprobe gefunden. Das sind 624 weniger als gewünscht. Offensichtlich gibt es nicht für jeden Satz genügend Material im Korpus, das mit den $k = 3$ Anfragetoken gefunden werden kann. Tabelle 3.11 zeigt, wie viele Matches je Anfragesatz gefunden wurden. 17 mal wurde gar kein Fuzzy-Match gefunden. Im Englischen steht deutlich häufiger ausreichend Material zur Verfügung als im Deutschen. Ob dies an der Art der Ermittlung der Kandidatenmenge liegt oder die tatsächliche Belegsituation wiederspiegelt, kann an dieser Stelle nicht beantwortet werden.

Eignung des Ähnlichkeitsmaßes

Um zu beurteilen, wie gut das Ähnlichkeitsmaß die Relevanz der Fuzzy-Matches vorhersagt, werden nun die Klassenhäufigkeiten in Abhängigkeit von den Ähnlichkeitswerten betrachtet. Eine Korrelationsanalyse wird hier nicht durchgeführt, da der Einarbeitungsaufwand hoch ist und dem Autor Erfahrungen fehlen, wie die Resultate zu interpretieren sind. Stattdessen werden die Ähnlichkeitswerte in Intervalle unterteilt und für jedes Inter-

5 Unterteilungen

$a\backslash$ Klasse*	0	20	60	70	75	80	85	90	95	100
$]0.2700, 1.0000]$	37	45	5	25	1	5	8	5	2	4
$]0.2377, 0.2700]$	17	9	0	1	0	0	0	0	0	0
$]0.2079, 0.2377]$	38	7	0	0	0	0	0	0	0	0
$]0.1675, 0.2079]$	22	6	0	0	0	0	0	0	0	0
$]0.0000, 0.1675]$	33	4	0	0	0	0	0	0	0	0

die ersten 10 von 100 Unterteilungen

$a\backslash$ Klasse*	0	20	60	70	75	80	85	90	95	100
$]0.6087, 1.0000]$	0	0	0	2	0	3	4	4	2	4
$]0.5147, 0.6087]$	0	0	2	11	0	1	3	1	0	0
$]0.4135, 0.5147]$	1	3	1	6	1	0	1	0	0	0
$]0.3801, 0.4135]$	1	5	0	4	0	0	0	0	0	0
$]0.3546, 0.3801]$	1	5	1	1	0	1	0	0	0	0
$]0.3401, 0.3546]$	0	7	0	0	0	0	0	0	0	0
$]0.3293, 0.3401]$	7	3	1	1	0	0	0	0	0	0
$]0.3207, 0.3293]$	2	4	0	0	0	0	0	0	0	0
$]0.3098, 0.3207]$	3	1	0	0	0	0	0	0	0	0
$]0.3000, 0.3098]$	3	3	0	0	0	0	0	0	0	0

* angegeben durch die Relevanz in %

Tabelle 3.12: Klassenverteilung in Ähnlichkeitsintervallen (Deutsch)

vall die absoluten Häufigkeiten der Klassen der Fuzzy-Matches mit Ähnlichkeitswerten aus dem Intervall ermittelt. Da die Verteilung der Ähnlichkeitswerte sprachabhängig ist, wird die Abhängigkeit für Deutsch und Englisch getrennt untersucht.

Der Idealfall wäre, dass solche Intervalle gefunden werden können, dass Intervalle und Klassen bijektiv und ordnungserhaltend einander zugeordnet sind. Das Ähnlichkeitsmaß würde dann auf den klassifizierten Fuzzy-Matches keine Vorhersagefehler machen, und man könnte eine sehr gute Vorhersagequalität bei neuen Fuzzy-Matches erwarten. (Oder es würde der Verdacht aufkommen, der Annotator habe die Sätze nicht gelesen und nur den Ähnlichkeitswerten Beachtung geschenkt.) Zu erwarten ist jedoch, dass in jedem Intervall mehrere Klassen vertreten sind, außer wenn man sie so schmal wählt, dass nur noch sehr wenige Fuzzy-Matches vertreten sind.

Die Tabellen 3.12 und 3.13 listen die Klassenhäufigkeiten für einige Intervalle auf. Die Klassen sind stellvertretend mit den in Tabelle 3.10 eindeutig zugeordneten Relevanzwerten angegeben, um Platz zu sparen. Die Intervalle wurden mit Hilfe der Rangliste aller Ähnlichkeitswerte der 2 035 deutschen bzw. 2 441 englischen Fuzzy-Matches so festgelegt, dass je Intervall möglichst gleich viele Fuzzy-Matches auftreten.[50]

Es wurden zwei verschiedene Unterteilungen vorgenommen. Die erste, nur fünf Intervalle umfassende Unterteilung zeigt, dass bei Ähnlichkeitswerten $a < 0,25$ nur sehr selten

[50] Die Zeilensummen in den Tabellen schwanken stark, da nur klassifizierte Fuzzy-Matches gezählt wurden und die Intervalleinteilung jedoch auf der Verteilung aller Matches beruht. Dies mag man bedauern, ist aber notwendig, um unabhängige und abhängige Größen in der Darstellung klar voneinander zu trennen.

5 Unterteilungen

$a\backslash$ Klasse*	0	20	60	70	75	80	85	90	95	100
$]0.2593, 1.0000]$	27	62	23	27	0	6	12	0	0	4
$]0.2227, 0.2593]$	21	10	0	1	0	0	0	0	0	0
$]0.1988, 0.2227]$	26	7	0	2	0	0	0	0	0	0
$]0.1675, 0.1988]$	30	11	0	0	0	0	0	0	0	0
$]0.0000, 0.1675]$	19	5	0	0	0	0	0	0	0	0

die ersten 10 von 100 Unterteilungen

$a\backslash$ Klasse*	0	20	60	70	75	80	85	90	95	100
$]0.6112, 1.0000]$	0	0	0	4	0	1	12	0	0	4
$]0.4363, 0.6112]$	0	6	3	6	0	1	0	0	0	0
$]0.3971, 0.4363]$	0	9	5	4	0	1	0	0	0	0
$]0.3693, 0.3971]$	1	6	2	4	0	1	0	0	0	0
$]0.3395, 0.3693]$	2	8	1	3	0	0	0	0	0	0
$]0.3234, 0.3395]$	1	4	0	1	0	1	0	0	0	0
$]0.3108, 0.3234]$	2	0	1	0	0	1	0	0	0	0
$]0.3031, 0.3108]$	2	3	3	1	0	0	0	0	0	0
$]0.2948, 0.3031]$	2	3	1	1	0	0	0	0	0	0
$]0.2906, 0.2948]$	2	1	1	0	0	0	0	0	0	0

* angegeben durch die Relevanz in %

Tabelle 3.13: Klassenverteilung in Ähnlichkeitsintervallen (Englisch)

bessere Fuzzy-Matches als Term-Matches auftreten. Unter 0,20 treten nur noch Term-Matches und irrelevante Matches auf, wobei Letztere die Mehrheit bilden. Im ersten Intervall dieser Unterteilung treten alle Klassen auf. Deshalb wurde noch eine weitere, feinere Unterteilung in 100 Intervalle vorgenommen. Die ersten zehn Intervalle, die der ersten Hälfte des ersten Intervalls der ersten Unterteilung entsprechen, sind in dem jeweils zweiten Teil der Tabellen abgebildet. Hier zeigt sich, dass sich die Häufigkeitsverteilung der Klassen mit steigendem Ähnlichkeitswert verbreitert und zu den Klassen höherer Relevanz verschiebt. Bei Werten über 0,61 treten nur noch Matches mit mindestens ähnlichem Inhalt auf.

Wichtig für die Entscheidung, den Schwerpunkt der Klassifikationsarbeit auf die Sätze der Stichprobe zu legen, für die die besten vier Fuzzy-Matches einen hohen Ähnlichkeitswertdurchschnitt aufweisen, ist auch, dass bei Ähnlichkeitswerten unter 0,4 nur noch sehr selten bessere Matches als Subsegment-Matches gefunden werden und dass unter 0,3 auch diese in der Regel ausbleiben.

Relevanz der Fuzzy-Matches

Klassifiziert wurden 567 Paare von Anfragesätzen und Fuzzy-Matches.[51] Das sind 12,7 % aller Fuzzy-Matches. Vorzugsweise wurden solche mit hoher Ähnlichkeit laut dem Ähnlichkeitsmaß klassifiziert. Tabelle 3.14 zeigt die Häufigkeiten der Klassen für die beiden

[51] Es sind nur 566 unterschiedliche Sätze. Einer ist Fuzzy-Match zu zwei verschiedenen Anfragesätzen.

Klasse	Relevanz	Häufigkeit	Deutsch	Englisch
Exact-Match	100 %	8	4	4
nur Tippfehler	95 %	2	2	0
gleicher Inhalt	90 %	5	5	0
fast gleicher Inhalt	85 %	20	8	12
enthält etwas mehr	80 %	11	5	6
enthält etwas weniger	75 %	1	1	0
ähnlicher Inhalt	70 %	56	26	30
Subsegment-Match	60 %	28	5	23
Term-Match	20 %	166	71	95
keine Relevanz	0 %	270	147	123
Summe		567	274	293

Tabelle 3.14: Häufigkeiten der Klassen

Sprachen und insgesamt.[52] 103 Fuzzy-Matches haben über 60 % Relevanz, also ähnlichen oder sogar fast gleichen Inhalt wie der Anfragesatz. Subsegment-Matches sind mehr als dreieinhalb mal seltener als relevantere Matches, Term-Matches wesentlich häufiger. Unterschiede zwischen den Sprachen können festgestellt werden: Im Englischen wurden keine Matches mit gleichem Inhalt, dafür jedoch entsprechend mehr mit fast gleichem Inhalt gefunden. Subsegment-Matches werden mehr als vier mal so häufig im Englischen als im Deutschen gefunden.

Sätze mit guter Beleglage

Es wurden zu 62 Sätzen der Stichprobe (31 je Sprache) Klassifikationen vorgenommen. Zu 17 weitere Anfragesätze gibt es keine Daten, da für sie gar keine Fuzzy-Matches im Korpus gefunden wurden. Betroffen sind 15 deutsche und 2 englische Sätze. Die Beleglage für die Anfragesätze lässt sich mit der Tabelle 3.14 schlecht einschätzen, da aus ihr nicht hervorgeht, ob sich die Fuzzy-Matches mit hoher Relevanz auf einige wenige Anfragesätze zurückgehen oder über viele verteilt sind. Dies ändert sich, wenn man für jede Klasse auszählt, wie häufig sie den besten Fuzzy-Match eines Anfragesatzes stellt. Der Tabelle 3.15 können die Häufigkeiten entnommen werden. Selbst wenn man annimmt, die nicht klassifizierten Matches hätten keine Relevanz, haben 36 von 510 Sätze der Stichprobe mindestens einen Fuzzy-Match mit Relevanz über 60 %. Das sind immerhin 7 %, im Deutschen etwas weniger (6,4 %) und im Englischen etwas mehr (7,7 %).

Zur Beurteilung der Belegsituation können auch die Tabellen im Anhang A.1 beitragen, die zu jedem Satz der Stichprobe die Güte der Fuzzy-Matches angeben, die als Durchschnitt der Relevanzwerte der jeweils besten vier Matches definiert wurde.

3.4.5 Bewertung

Es wurde genügend Material für die Diskussion im nachfolgenden Kapitel gefunden. Lediglich die Fuzzy-Match-Klassen „nur Tippfehler“ und „enthält etwas mehr“ nicht sehr schwach vertreten. Im Englischen gibt es darüber hinaus keine Beispiele für Matches der

[52] Für die einzelnen Sprachen sind die Werte die Spaltensummen der Tabellen 3.12 und 3.13.

Klasse	Relevanz	Häufigkeit	Deutsch	Englisch
Exact-Match	100 %	7	3	4
nur Tippfehler	95 %	0	0	0
gleicher Inhalt	90 %	3	3	0
fast gleicher Inhalt	85 %	7	4	3
enthält etwas mehr	80 %	4	1	3
enthält etwas weniger	75 %	0	0	0
ähnlicher Inhalt	70 %	15	5	10
Subsegment-Match	60 %	5	2	3
Term-Match	20 %	12	9	3
keine Relevanz	0 %	9	4	5
Summe		62	31	31

Tabelle 3.15: Klassenhäufigkeiten bei den besten Fuzzy-Matches

Klasse „gleicher Inhalt". Es gibt 36 Sätze mit Fuzzy-Matches hoher Relevanz und 5 Sätze, zu denen immerhin noch (mindestens) ein Subsegment-Match gefunden wird.

Im Anhang A.2 werden die fünf deutschen und sechs englischen Sätze der Stichprobe, die die bester Belegsituation (gemäß der eingeführten Güte) aufweisen, mit ihren Fuzzy-Matches aufgelistet.

3.5 Zusammenfassung

Das KoKS-Korpus ist ein paralleles Korpus mit den Sprachen Deutsch und Englisch und umfasst je Sprache etwa viereinhalb Millionen Wörter. Es ist mit POS-Tags und Lemmata annotiert und auf Satzebene alignt. Über die KoKS-Datenbank kann das Korpus flexibel eingesetzt werden. Indizes ermöglichen einen schnellen Zugriff auf das Korpus.

Im Rahmen dieser Arbeit wurden weitere Indizes implementiert, um die Fuzzy-Match-Suche effizienter durchführen zu können. Dabei wurde auch eine Lösung zu dem Problem entwickelt, dass im KoKS-Korpus die Grundformalternativen nicht explizit repräsentiert, sondern als ein mit einem speziellen Zeichen separierter String gespeichert werden, der in dieser Form vom IMS TreeTagger annotiert wird.

Es wurden Aspekte der Vorverarbeitung beleuchtet, die im KoKS Abschlussbericht nicht behandelt werden. Insbesondere wurde die Funktionsweise des IMS TreeTaggers erklärt, Probleme der Segmentierung aufgezeigt und auf Schwächen des KoKS-Aligners hingewiesen.

Mit der Aufnahme des Harry Potter Teilkorpus ist eine neue Textsorte im Korpus vertreten, die andere Eigenschaften hat, als die anderen Teilkorpora. Dies sind die wörtliche Rede und das sehr häufige Auftreten eines bestimmten Eigennamens. Trotzdem dominiert das EU Teilkorpus das Korpus durch seine Größe.

3.5.1 Ausblick

Das KoKS-Korpus kann verbessert werden, indem die Vollformliste für die Umlaut- und Eszettkorrektur sorgfältiger aufgebaut wird. Dazu müssen die Wörterbüch und Teilkorpora

nochmal geprüft und eine neue Reihenfolge für das Eintragen in die Datenbank gewählt werden.

Kapitel 4

Bilinguale Korpora in CAT-Systemen - eine Anwendungsperspektive

Die Anwendung eines Translation Memorys stößt an ihre Grenzen, wenn nur Subsegment-Matches oder Fuzzy-Matches mit geringer inhaltlicher Ähnlichkeit gefunden werden. Werkzeuge, mit denen ein CAT-System den Übersetzer auch in diesen Situationen unterstützen kann, wurden im Abschnitt 1.1 vorgestellt. Sowohl datengestützt als auch automatisch erstellt neben einem Translation Memory nur die EBMT (Example-Based Machine Translation) Übersetzungsvorschläge.

EBMT-Ansätze unterscheiden sich sehr in der Art des linguistischen Wissens, das sie einsetzen. Ein Teil der Ansätze extrahiert Transferregeln aus den Daten, die dann in einem klassischen MT-System verwendet werden. McTait (2001) bildet in einer Vorverarbeitungsphase flache Strukturen, die Variablen enthalten, um sie in der Übersetzungsphase auf die zu übersetzenden Sätze anzuwenden. Die Extraktion der Strukturen, die er Translation Patterns nennt, erfolgt mit einem machinellen Lernverfahren. Linguistisches Wissen in Form von POS-Tags oder Grundformannotationen wird nicht eingesetzt. Einführungen und Übersichten zu EBMT bieten Carl und Way (2003) und Somers (1999).

In Nachfolgendem wird eine Anwendungsperspektive gezeigt, die keinen dieser Ansätze verfolgt, sondern sich im wesentlichen auf Alignment und Ähnlichkeitsmaße stützt.

4.1 Ein Ansatz zur Nutzung mehrerer TUs

Der in diesem Abschnitt skizzierte Ansatz nutzt mehrere Translation Units des Refenrenzmaterial, die mindestens ein Subsegment mit dem zu übersetzenden Satz gemeinsam haben. Kennzeichnet für diesen Ansatz ist, dass er für alle drei grundlegenden Schritte Subsegment-Suche, Identifikation der Übersetzungen und Kombination derselben Alignment-Techniken nutzt. Dies ermöglicht es, je nach Verfügbarkeit unterschiedlich viel linguistisches Wissen einzusetzen.

4.1.1 Subsegment-Suche

Im Abschnitt **??** wurde festgestellt, dass Subsegment-Matches wesentlich seltener als Matches mit ähnlichem Inhalt auftreten. Man könnte daher fragen, ob der Aufwand für die Generierung eines Übersetzungsvorschlags aus Subsegment-Matches gerechtfertigt ist. Zwei Punkte sprechen dafüt: Zum einen muss man bedenken, dass die Unterscheidung zwischen Subsegment-Matches und Matches mit ähnlichem Inhalt manuell getroffen wurde. In der Praxis steht nur das Ähnlichkeitsmaß zur Verfügung. Es hat sich aber gezeigt, dass das Maß die Klassen nicht eindeutig bestimmen kann. In einem breiten Intervall von Ähnlichkeitswerten treten im Englischen sowohl Subsegment-Matches als auch Matches mit ähnlichem Inhat auf. Im Deutschen gibt es Überschneidungen mit beiden Nachbarklassen „Term-Match" und „Match mit ähnlichem Inhalt". Die Ähnlichkeitsschwelle, ab der wie in einem klassischen Translation Memory ein Übersetzungsvorschlag aus einer Fundstelle übernommen wird, sollte also so hoch gewählt werden, dass keine (oder nur wenige) Subsegment-Matches als Match mit ähnlichem Inhalt dem Übersetzer präsentiert wird. Ebenso ist denkbar, Übersetzungen aus Fuzzy-Matches und generierte Übersetzungen dem Benutzer zugleich anzubieten.

Als zweites kann die Mindestlänge der Subsegmente herabgesetzt werden. Bei der Klassifikation wurden acht Token verlangt. Kürzere Subsegmente, wie z. B. „die Voraussetzungen von Artikel 66" (Segment 612370-de)[1], „der zweiten Stufe des Vertragsverletzungsverfahrens" (Segment 457666-de) und „sent a reasoned opinion to" (Segment 457677-en), können zugelassen werden. Die große Zahl von bei der Klassifikation gefundenen Term-Matches deutet an, dass mit kleineren Subsegmentlängen die Zahl der Matches vervielfacht werden kann.

Im Folgenden werden auch Fuzzy-Matches mit ähnlichem bis fast gleichem Inhalt verwendet. Auf eine feste Längenbeschränkung der Subsegmente wird verzichtet.

Hilfsmittel

Die tabellarische Darstellung eines sequentiellen Wortalignments im Annotationstool erlaubt es, übereinstimmende Subsegmente schnell zu erkennen. Allerdings verlangt das Tool eine genaue zeichenweise Übereinstimmung. Beispielsweise führt beim Match 612370-de-632624 (Seite 89) ein Fehler bei der Eszett-Korrektur zur Aufteilung des Subsegments „dieser Übernahme die Voraussetzungen des Artikel 66". Die Abweichung führt dazu, dass das betroffene Wort alleine in einer Zeile steht. Solche Subsegmente können daher auch leicht erkannt werden. Im Folgenden werden sie besonders gekennzeichent, wenn sie verwendet werden.

Beispiele für Subsegmente

Zu zwei Segment zeigt Tabelle 4.1 Subsegmente, die nicht Teil eines größeren Subsegments sind, d. h. maximal sind. Die Anfragesätze sind:

(4.1) Da mit dieser übernahme die Voraußetzungen von Artikel 66 Absatz 2 EGKS-Vertrag erfüllt sind , hat die Kommission ihre Zustimmung erteilt

(4.2) Die Aufforderungen ergehen jeweils in Form einer mit Gründen versehenen Stellungnahme # , der zweiten Stufe des Vertragsverletzungsverfahrens gemäß Artikel 226 EG-Vertrag .

[1] Siehe Anhang A.2. Fuzzy-Matches werden im Folgenden als Tupel von Anfragesegment, Sprache und Referenzsegment angegeben.

Anfrage	Match	Subsegment
612370-de	619902	die Voraußetzungen von Artikel 66 Absatz 2 EGKS-Vertrag erfüllt sind
612370-de	613006	ihre Zustimmung erteilt
612370-de	625456	, hat die Kommission ihre
457666-de	464651	mit Gründen versehenen Stellungnahme #, der zweiten Stufe des Vertragsverletzungsverfahrens
457666-de	470928	Artikel 226 EG-Vertrag
ähnliche Subsegmente		
612370-de	613006	die Kommission hat deshalb ihre Zustimmung erteilt
612370-de	625456	, hat die Kommission ihre Genehmigung erteilt
457666-de	435175	Die Aufforderung an Belgien ergeht

Tabelle 4.1: Beispiele für Subsegmente (Deutsch)

In der Tabelle wurden auch mehrere Subsegmente mit ähnlichem Inhalt aufgeführt. In einem Fall wie 457666-de-435175 könnte das Trigramm-Ähnlichkeitsmaß benutzt werden, um die Ähnlichkeit festzustellen. Für kurze Subsegmente ist dies nicht möglich, da kleine Änderungen bereits zu einem großen Anteil von Trigrammen führen, die nicht in beiden Subsegmenten gleich häufig vorkommen. Auch in dem Beispiel, in dem „Zustimmung“ durch „Genehmigung“ ausgetauscht ist, ist das Maß ungeeignet, da sich sehr viele Trigramm-Häufigkeiten durch die Substitution ändern.

Um die Ähnlichkeitsschwelle weiter absenken zu können ohne viele irrelevante Subsegmente zu finden, sind weitere Kriterien erforderlich, die ein Subsegment erfüllen muss. Beispielsweise könnte man verlangen, dass das Subsegment die gleiche POS-Tagfolge aufweisen muss, wie das entsprechende Subsegment im Anfragesatz. Im Fall 612370-de-625456 würde dies die Änderung von „Zustimmung“ zu „Genehmigung“ erlauben, den Einschub „hat deshalb“ in 612370-de-613006 dagegen verbieten. Entwickelt man diese Kriterien weiter, stellt man schließlich fest, dass ein monolingualer Subsegment-Aligner vorliegt.

Wenn die erste Fuzzy-Match-Anfrage mit dem Ausgangssatz nicht genug Material aus dem Korpus extrahiert hat, könnte man neue Anfragen mit den bereits identifizierten Subsegmenten und mit den noch nicht abgedeckten Textfragmenten starten.

4.1.2 Identifikation der Übersetzung eines Subsegments

Naheliegend wäre es, ein Wortalignment zu erstellen und dann als Übersetzung eines Subsegments die zugeordneten Wörter zu verwendet. Dies erfordert aber mehr Aufwand als nötig. Soll z. B. das Subsegment „die Kommission hat deshalb ihre Zustimmung erteilt“ im Segment 613006 (Seite 89) übersetzt werden, dann ist es irrelevant, wie die einzelnen Wörter aus dem Subsegment und im vorangehenden Satzteil übersetzt sind.

Zum Alignen kann fest vorgeben, dass je Sprache nur zwei Gruppen vorhanden sind, nämlich die Token, die zum Subsegment gehören, und alle übrigen Token. Im ausgangs-

Subsegment	Übersetzung
die Voraußetzungen von Artikel 66 Absatz 2 EGKS-Vertrag erfüllt sind	it meets the test of authorisation in Article 66 (2) ECSC Treaty
ihre Zustimmung erteilt	granted its authorisation
, hat die Kommission ihre	The Commission has
mit Gründen versehenen Stellungnahme #, der zweiten Stufe des Vertragsverletzungsverfahrens	# reasoned opinion # , the second stage of infringement proceedings
Artikel 226 EG-Vertrag	Article 226 of the EC Treaty
ähnliche Subsegmente	
die Kommission hat deshalb ihre Zustimmung erteilt	the Commission has therefore granted its authorisation
, hat die Kommission ihre Genehmigung erteilt	The Commission has authorized
Die Aufforderung an Belgien ergeht	The request will be made

Tabelle 4.2: Übersetzungen der Subsegmente

sprachlichen Satz ist die Zugehörigkeit bereits festgelegt. Der Aligner muss nur die beste Zuordnung der zielsprachlichen Token zu den beiden Alignment-Beads bestimmen. Wenn man annimmt, dass die Übersetzung des Subsegments wieder zusammenhängend ist, dann sind nur zwei Parameter, Start und Ende des Subsegments in der Übersetzung, zu bestimmen.

Tabelle 4.2 zeigt manuell identifizierte Übersetzungen der Subsegmente aus dem Beispiel. Zum Subsegment „, hat die Kommission ihre“ ist die Auswahl eines zielsprachlichen Subsegments als Übersetzung schwierig, da es keine vollständige Phrase beinhaltet.

4.1.3 Direkte Zuordnung möglicher Übersetzungen

Soweit der Ansatz beschrieben ist, werden zu einzelnen Subsegmenten des Anfragesatzes Übersetzungen aus dem Referenzmaterial extrahiert. Dabei werden zwei Subsegment-Alignments eingesetzt. Zuerst werden monolinguale Entsprechnungen zwischen Anfragesatz und Fundstelle gesucht. Zu den so gefundenen Subsegmente werden dann durch das bilinguale Alignment Übersetzungen ermittelt.

Alternativ könnte man die Übersetzungen der Fuzzy-Matches direkt mit dem Anfragesatz alignen. Dabei müsste das Optimierungsziel angepasst werden. Es sollen einzelne sehr gute Alignment Beads gefunden werden. Das Gesamtalignment darf schlecht sein. Die Alignment Beads mit guter Zuordnungsqualität liefern dann die Subsegmente.

Die zweistufige Lösung überlässt die Identifikation der nutzbaren Teile des Fuzzy-Matches dem monolingualen Alignment. Hier können strenge Kriterien, wie z. B. das gemeinsame Auftreten aller Wörter, angewendet werden. Die einstufige Lösung muss sich hier darauf verlassen, dass das bilinguale Alignment keine falschen Zuordnungen enthält.

4.1.4 Generierung des Übersetzungsvorschlags

Eine Möglichkeit für die Generierung des Übersetzungsvorschlags wäre, Subsegmente zur Abdeckung eines möglichst großen Teils des Anfragesatzes auszuwählen und deren Übersetzung einfach aneinander zu hängen. Es wäre dann Aufgabe des Übersetzers, die Fragmente richtig zu ordnen und anzupassen.

Auch hier könnte man Alignment-Techniken nutzen, um die Reihenfolge der Übersetzungsfragmente zu dem Anfragesatz passt. Im Falle von nur zusammenhängenden Subsegmenten ist dies trivial, da sie dann nur nach ihrer Startposition im Anfragesatz sortiert werden müssen. Für diskontinuierliche Subsegmente könnte ein Aligner verschiedene Anordnungen bewerten.

Zusätzlich zur Plausibilität des Alignments zwischen Ausgangssatz und Übersetzungsvorschlag könnte noch die zielsprachliche Plausibilität bewertet werden. Flache Analysestrukturen bieten hierfür Hilfsmittel. Beispielsweise können die POS-Tagfolgen daraufhin geprüft werden, ob sie im Korpus belegt sind.

Grundsätzlich neu sind diese Vorschläge nicht. Beispielsweise nutzen Somers et al. (1994) POS-Tagfolgen des Kontexts eines Fragments, die im Referenzmaterial belegt sind, um aus Kombinationsalternativen die plausibelste auszuwählen (Seite 8).

4.2 Zusammenfassung

Im Rahmen dieser Arbeit wurde dargestellt, wie ein großes, bilinguales Korpus für die datengestützte Übersetzung nutzbar gemacht werden kann. Das KoKS-Sytem wurde entsprechend erweitert. Die Ermittlung von Fuzzy-Matches nahm einen großen Raum ein. Sie ist Grundlage für den in diesem Kapitel skizzierten Ansatz zur automatischen Übersetzung mit flachen Analysestrukturen.

Anhang A

Fuzzy-Matches

A.1 Stichprobe

Die nachfolgenden Tabellen zeigen Daten zu den Sätzen (genauer: Segmenten) der Stichprobe geordnet nach der durchschnittlichen Ähnlichkeit der besten vier Fuzzy-Matches. Angegeben sind zusätzlich die Längen der Sätze in Token sowie die Güte der Belegsituation, die aus der Klassifikation der Fuzzy-Matches berechnet wird. Fehlt letzterer Wert, dann wurden die Fuzzy-Matches zu dem betroffenen Satz (noch) nicht klassifiziert.

A.1.1 Deutsche Sätze der Stichprobe

250 Sätze der deutschen Korpushälfte wurden ausgewählt.

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
612370	21	93.1 %	0.4 %	98 %
457666	23	71.4 %	0.8 %	86 %
464698	29	62.1 %	1.2 %	85 %
616161	42	55.0 %	1.6 %	81 %
652804	51	53.0 %	2.0 %	62 %
448876	39	48.8 %	2.4 %	51 %
478762	16	46.0 %	2.8 %	36 %
444774	37	45.7 %	3.2 %	40 %
642256	43	43.7 %	3.6 %	49 %
520954	14	43.4 %	4.0 %	30 %
631122	27	41.9 %	4.4 %	32 %
482278	19	39.5 %	4.8 %	30 %
455908	31	37.9 %	5.2 %	-
621160	38	37.5 %	5.6 %	-
687378	12	37.3 %	6.0 %	-
503374	17	37.0 %	6.4 %	35 %
632880	13	36.1 %	6.8 %	-
456494	31	35.6 %	7.2 %	-

Fortsetzung nächste Seite

Übersicht zur Stichprobe (Deutsch)

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
475246	17	35.4 %	7.6 %	-
681518	13	35.1 %	8.0 %	0 %
505718	13	34.9 %	8.4 %	-
617644	26	34.8 %	8.8 %	-
484622	18	34.7 %	9.2 %	-
471730	24	34.1 %	9.6 %	-
655734	36	34.1 %	10.0 %	35 %
468800	32	33.4 %	10.4 %	-
636396	44	33.1 %	10.8 %	-
490482	12	33.1 %	11.2 %	-
465870	17	32.7 %	11.6 %	32 %
613542	22	32.5 %	12.0 %	-
634638	30	32.4 %	12.4 %	-
474660	32	32.4 %	12.8 %	-
641084	28	32.3 %	13.2 %	-
431882	38	32.3 %	13.6 %	32 %
670970	14	32.1 %	14.0 %	-
615300	40	32.1 %	14.4 %	-
644014	28	32.1 %	14.8 %	-
421920	56	31.2 %	15.2 %	10 %
637568	20	30.4 %	15.6 %	-
648702	43	29.7 %	16.0 %	-
451806	28	29.7 %	16.4 %	-
443016	17	29.0 %	16.8 %	35 %
615886	43	28.9 %	17.2 %	-
431296	37	28.9 %	17.6 %	-
645502	23	28.8 %	18.0 %	-
516266	13	28.7 %	18.4 %	-
428952	54	28.5 %	18.8 %	-
635224	31	28.5 %	19.2 %	-
628778	32	28.3 %	19.6 %	-
434226	60	28.2 %	20.0 %	10 %
445360	13	28.1 %	20.4 %	-
446532	34	28.1 %	20.8 %	-
657492	40	28.1 %	21.2 %	-
447118	21	28.1 %	21.6 %	-
427780	51	28.0 %	22.0 %	-
441844	43	27.8 %	22.4 %	-
652218	38	27.7 %	22.8 %	-
649288	22	27.7 %	23.2 %	-
433640	59	27.6 %	23.6 %	-
450048	44	27.5 %	24.0 %	-
496928	22	27.5 %	24.4 %	-
498686	16	27.4 %	24.8 %	-
424264	55	27.2 %	25.2 %	-

Fortsetzung nächste Seite

Übersicht zur Stichprobe (Deutsch)

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
469972	44	27.1 %	25.6 %	-
461182	17	26.9 %	26.0 %	-
430710	20	26.9 %	26.4 %	-
436570	31	26.9 %	26.8 %	-
437742	42	26.9 %	27.2 %	-
489896	18	26.8 %	27.6 %	-
425436	56	26.8 %	28.0 %	-
437156	30	26.6 %	28.4 %	-
458838	21	26.6 %	28.8 %	-
424850	24	26.6 %	29.2 %	-
520368	17	26.5 %	29.6 %	-
679174	12	26.5 %	30.0 %	18 %
639912	41	26.4 %	30.4 %	-
673900	13	26.3 %	30.8 %	-
633466	46	26.3 %	31.2 %	-
491068	51	26.2 %	31.6 %	-
678002	45	26.2 %	32.0 %	-
473488	44	26.2 %	32.4 %	-
638740	31	26.1 %	32.8 %	-
649874	52	26.1 %	33.2 %	-
501030	19	26.1 %	33.6 %	-
458252	57	26.1 %	34.0 %	-
629364	36	26.0 %	34.4 %	-
442430	39	25.9 %	34.8 %	-
620574	25	25.7 %	35.2 %	-
647530	32	25.7 %	35.6 %	-
503960	15	25.6 %	36.0 %	-
624676	41	25.4 %	36.4 %	-
441258	31	25.4 %	36.8 %	-
654562	32	25.4 %	37.2 %	-
634052	37	25.3 %	37.6 %	-
524470	25	25.2 %	38.0 %	-
427194	60	25.2 %	38.4 %	-
653976	26	25.2 %	38.8 %	-
618816	43	25.1 %	39.2 %	-
645186	37	25.1 %	39.6 %	-
640498	15	25.0 %	40.0 %	5 %
639326	33	24.9 %	40.4 %	-
638154	19	24.9 %	40.8 %	-
434812	21	24.8 %	41.2 %	-
492240	19	24.8 %	41.6 %	-
618230	33	24.7 %	42.0 %	-
426608	38	24.7 %	42.4 %	-
440115	28	24.6 %	42.8 %	-
619402	32	24.5 %	43.2 %	-

Fortsetzung nächste Seite

Übersicht zur Stichprobe (Deutsch)

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
508062	18	24.5 %	43.6 %	-
636982	31	24.5 %	44.0 %	-
428366	33	24.3 %	44.4 %	-
651046	45	24.3 %	44.8 %	-
516852	22	23.9 %	45.2 %	-
448290	49	23.8 %	45.6 %	-
627020	14	23.8 %	46.0 %	-
472902	27	23.7 %	46.4 %	-
664524	35	23.5 %	46.8 %	-
486380	21	23.5 %	47.2 %	-
651632	36	23.4 %	47.6 %	-
658078	32	23.4 %	48.0 %	-
460010	24	23.4 %	48.4 %	-
435398	16	23.4 %	48.8 %	-
435984	23	23.2 %	49.2 %	-
628192	22	23.2 %	49.6 %	-
464112	34	23.1 %	50.0 %	5 %
449462	15	23.1 %	50.4 %	-
650460	58	23.1 %	50.8 %	-
614128	37	22.9 %	51.2 %	-
440672	31	22.8 %	51.6 %	-
627606	27	22.8 %	52.0 %	-
494584	29	22.7 %	52.4 %	-
469386	18	22.6 %	52.8 %	-
622332	38	22.3 %	53.2 %	-
682690	18	22.2 %	53.6 %	-
530330	20	22.2 %	54.0 %	-
631708	20	22.2 %	54.4 %	-
422506	39	22.2 %	54.8 %	-
501616	15	22.2 %	55.2 %	-
661594	16	22.1 %	55.6 %	-
663938	43	22.0 %	56.0 %	-
672728	18	22.0 %	56.4 %	-
500444	19	21.9 %	56.8 %	-
426022	23	21.9 %	57.2 %	-
477590	24	21.8 %	57.6 %	-
658664	33	21.7 %	58.0 %	-
499272	40	21.7 %	58.4 %	-
655148	46	21.7 %	58.8 %	-
635810	20	21.6 %	59.2 %	-
523298	15	21.6 %	59.6 %	-
438328	20	21.6 %	60.0 %	0 %
447704	33	21.5 %	60.4 %	-
460596	36	21.3 %	60.8 %	-
482864	14	21.3 %	61.2 %	-

Fortsetzung nächste Seite

Übersicht zur Stichprobe (Deutsch)

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
623504	18	21.2 %	61.6 %	-
526814	38	21.2 %	62.0 %	-
451220	32	21.2 %	62.4 %	-
523884	41	21.1 %	62.8 %	-
454736	22	20.9 %	63.2 %	-
509820	16	20.8 %	63.6 %	-
624090	38	20.8 %	64.0 %	-
512164	22	20.8 %	64.4 %	-
488138	15	20.7 %	64.8 %	-
528572	19	20.7 %	65.2 %	20 %
450634	25	20.7 %	65.6 %	-
625262	18	20.4 %	66.0 %	-
656906	22	20.4 %	66.4 %	-
459424	32	20.4 %	66.8 %	-
440086	33	20.2 %	67.2 %	-
468214	19	19.9 %	67.6 %	-
621746	24	19.8 %	68.0 %	-
643428	20	19.6 %	68.4 %	-
463526	14	19.5 %	68.8 %	-
472316	16	19.3 %	69.2 %	-
465284	32	19.3 %	69.6 %	-
423678	25	19.2 %	70.0 %	15 %
471144	24	19.0 %	70.4 %	-
515680	22	18.9 %	70.8 %	-
486966	13	18.8 %	71.2 %	-
632294	21	18.5 %	71.6 %	-
512750	24	18.5 %	72.0 %	-
662766	16	18.3 %	72.4 %	-
685034	26	18.1 %	72.8 %	-
527400	21	17.9 %	73.2 %	-
487552	19	17.8 %	73.6 %	-
646944	22	17.7 %	74.0 %	-
687964	13	17.7 %	74.4 %	-
510406	24	17.6 %	74.8 %	-
452978	23	17.6 %	75.2 %	10 %
453564	14	17.5 %	75.6 %	-
457080	22	17.5 %	76.0 %	-
648116	16	17.3 %	76.4 %	-
455322	31	16.9 %	76.8 %	-
467628	18	16.6 %	77.2 %	-
510992	17	16.5 %	77.6 %	-
470558	18	16.3 %	78.0 %	-
525056	24	16.3 %	78.4 %	-
443602	18	16.2 %	78.8 %	-
645772	19	16.2 %	79.2 %	-

Fortsetzung nächste Seite

Übersicht zur Stichprobe (Deutsch)

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
593032	12	16.2 %	79.6 %	-
525642	20	16.1 %	80.0 %	0 %
423092	22	15.9 %	80.4 %	-
622918	26	15.6 %	80.8 %	-
467042	15	15.6 %	81.2 %	-
526228	12	15.0 %	81.6 %	-
616472	17	14.7 %	82.0 %	-
522712	21	14.3 %	82.4 %	-
454150	13	14.1 %	82.8 %	-
502788	17	13.7 %	83.2 %	-
665110	21	13.6 %	83.6 %	-
444188	16	13.1 %	84.0 %	-
432468	27	12.4 %	84.4 %	-
518610	16	12.1 %	84.8 %	-
521540	12	11.2 %	85.2 %	-
493412	12	10.3 %	85.6 %	-
671556	24	9.8 %	86.0 %	-
626434	16	9.0 %	86.4 %	-
505132	17	8.9 %	86.8 %	-
662180	18	8.6 %	87.2 %	-
625848	18	8.6 %	87.6 %	-
498797	33	8.2 %	88.0 %	-
666282	21	8.0 %	88.4 %	-
498100	17	5.8 %	88.8 %	-
496342	15	5.6 %	89.2 %	-
666868	28	5.2 %	89.6 %	-
518024	20	4.6 %	90.0 %	5 %
672142	22	4.3 %	90.4 %	-
433054	22	4.2 %	90.8 %	-
682104	18	3.8 %	91.2 %	-
492826	13	3.7 %	91.6 %	-
469456	20	3.2 %	92.0 %	0 %
531502	16	3.1 %	92.4 %	-
529744	18	3.1 %	92.8 %	-
507476	30	2.9 %	93.2 %	-
656320	24	2.9 %	93.6 %	-
504546	16	2.0 %	94.0 %	5 %
445946	14	0.0 %	94.4 %	0 %
476418	12	0.0 %	94.8 %	0 %
479934	17	0.0 %	95.2 %	0 %
481692	20	0.0 %	95.6 %	0 %
483450	12	0.0 %	96.0 %	0 %
497514	17	0.0 %	96.4 %	0 %
509234	18	0.0 %	96.8 %	0 %
522126	12	0.0 %	97.2 %	0 %

Fortsetzung nächste Seite

Übersicht zur Stichprobe (Deutsch)

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
614714	13	0.0 %	97.6 %	0 %
642842	17	0.0 %	98.0 %	0 %
653390	21	0.0 %	98.4 %	0 %
669798	16	0.0 %	98.8 %	0 %
670384	16	0.0 %	99.2 %	0 %
675658	12	0.0 %	99.6 %	0 %
678588	18	0.0 %	100.0 %	0 %

Tabelle A.1: Übersicht zur Stichprobe (Deutsch)

A.1.2 Deutsche Sätze der Stichprobe

260 Sätze der englischen Korpushälfte wurden ausgewählt.

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
461295	18	92.0 %	0.4 %	85 %
656064	15	64.5 %	0.8 %	70 %
461898	39	54.0 %	1.2 %	62 %
620487	21	51.8 %	1.5 %	35 %
466722	23	51.8 %	1.9 %	74 %
648828	48	48.5 %	2.3 %	40 %
456471	22	46.8 %	2.7 %	32 %
449838	20	44.2 %	3.1 %	40 %
441396	25	42.9 %	3.5 %	70 %
625914	40	42.8 %	3.8 %	52 %
429939	28	41.4 %	4.2 %	72 %
655461	43	40.7 %	4.6 %	60 %
435366	22	40.3 %	5.0 %	21 %
641592	22	40.3 %	5.4 %	-
628929	30	40.0 %	5.8 %	-
621090	41	39.9 %	6.2 %	-
520992	12	39.5 %	6.5 %	25 %
448029	18	38.5 %	6.9 %	-
638577	16	38.4 %	7.3 %	-
471546	23	38.0 %	7.7 %	-
619884	50	37.5 %	8.1 %	57 %
472149	25	37.1 %	8.5 %	-
493254	25	37.0 %	8.8 %	-
450441	26	36.2 %	9.2 %	-
683802	13	35.8 %	9.6 %	-
649431	49	33.4 %	10.0 %	48 %
624708	43	33.2 %	10.4 %	-

Fortsetzung nächste Seite

Übersicht zur Stichprobe (Englisch)

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
690435	13	32.5 %	10.8 %	-
640989	35	32.4 %	11.2 %	-
423306	22	32.3 %	11.5 %	42 %
630135	38	32.2 %	11.9 %	-
616869	33	32.1 %	12.3 %	-
447426	54	31.8 %	12.7 %	-
523404	14	31.2 %	13.1 %	-
457677	21	31.0 %	13.5 %	75 %
689832	14	30.8 %	13.8 %	-
635562	29	30.7 %	14.2 %	-
438381	22	30.7 %	14.6 %	-
473958	26	30.5 %	15.0 %	20 %
468531	43	30.2 %	15.4 %	-
674154	25	30.0 %	15.8 %	-
432351	34	29.9 %	16.2 %	-
481194	20	29.7 %	16.5 %	-
657270	45	29.7 %	16.9 %	32 %
451647	40	29.6 %	17.3 %	-
668727	12	29.3 %	17.7 %	-
445617	42	29.1 %	18.1 %	-
613251	33	29.0 %	18.5 %	-
644004	25	28.9 %	18.8 %	-
659079	31	28.8 %	19.2 %	-
514962	28	28.7 %	19.6 %	-
658476	43	28.7 %	20.0 %	60 %
474561	60	28.6 %	20.4 %	-
443205	54	28.6 %	20.8 %	-
674757	12	28.3 %	21.2 %	-
650637	53	28.2 %	21.5 %	-
648225	30	28.1 %	21.9 %	-
675963	12	27.8 %	22.3 %	-
422703	45	27.8 %	22.7 %	-
631341	49	27.7 %	23.1 %	-
436572	20	27.6 %	23.5 %	-
507123	20	27.5 %	23.8 %	-
442602	54	27.4 %	24.2 %	-
425115	49	27.3 %	24.6 %	-
470340	32	27.1 %	25.0 %	-
521595	23	26.9 %	25.4 %	-
495666	21	26.8 %	25.8 %	-
514359	23	26.7 %	26.2 %	-
435969	53	26.5 %	26.5 %	-
505314	14	26.4 %	26.9 %	-
465516	21	26.3 %	27.3 %	-
484812	42	26.3 %	27.7 %	-

Fortsetzung nächste Seite

Übersicht zur Stichprobe (Englisch)

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
524610	19	26.3 %	28.1 %	-
432954	33	26.3 %	28.5 %	-
467325	46	26.3 %	28.8 %	-
652446	25	26.2 %	29.2 %	-
490239	13	26.1 %	29.6 %	-
637974	46	26.1 %	30.0 %	30 %
473355	43	26.0 %	30.4 %	-
624105	40	26.0 %	30.8 %	-
504108	37	26.0 %	31.2 %	-
673551	24	26.0 %	31.5 %	-
494460	23	25.8 %	31.9 %	-
448632	41	25.4 %	32.3 %	-
424512	13	25.4 %	32.7 %	-
622899	19	25.4 %	33.1 %	-
632547	48	25.4 %	33.5 %	-
462501	14	25.2 %	33.8 %	-
481797	33	24.9 %	34.2 %	-
645210	26	24.9 %	34.6 %	-
634959	24	24.8 %	35.0 %	-
634356	28	24.7 %	35.4 %	-
670536	34	24.6 %	35.8 %	-
657873	15	24.6 %	36.2 %	-
460692	59	24.6 %	36.5 %	-
488430	44	24.6 %	36.9 %	-
433557	31	24.4 %	37.3 %	-
517374	15	24.3 %	37.7 %	-
622296	47	24.1 %	38.1 %	-
464310	27	24.0 %	38.5 %	-
627723	17	24.0 %	38.8 %	-
654255	35	24.0 %	39.2 %	-
520389	25	23.9 %	39.6 %	-
618678	33	23.9 %	40.0 %	45 %
429336	28	23.8 %	40.4 %	-
659682	41	23.8 %	40.8 %	-
633753	25	23.8 %	41.2 %	-
504711	16	23.8 %	41.5 %	-
615060	24	23.7 %	41.9 %	-
680787	14	23.7 %	42.3 %	-
502902	29	23.5 %	42.7 %	-
486018	23	23.5 %	43.1 %	-
492651	19	23.4 %	43.5 %	-
441999	41	23.4 %	43.8 %	-
451044	29	23.3 %	44.2 %	-
653049	55	23.3 %	44.6 %	-
455868	19	23.3 %	45.0 %	-

Fortsetzung nächste Seite

Übersicht zur Stichprobe (Englisch)

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
643401	41	23.3 %	45.4 %	-
639180	31	23.2 %	45.8 %	-
508329	16	23.2 %	46.2 %	-
426321	16	23.2 %	46.5 %	-
650034	32	23.1 %	46.9 %	-
615663	23	23.0 %	47.3 %	-
645813	58	23.0 %	47.7 %	-
478782	24	23.0 %	48.1 %	-
515565	17	23.0 %	48.5 %	-
498681	16	23.0 %	48.8 %	-
512550	33	22.9 %	49.2 %	-
688023	34	22.9 %	49.6 %	-
430542	38	22.9 %	50.0 %	0 %
663300	18	22.9 %	50.4 %	-
642195	39	22.8 %	50.8 %	-
612648	17	22.7 %	51.2 %	-
633150	19	22.7 %	51.5 %	-
663903	14	22.6 %	51.9 %	-
522801	26	22.6 %	52.3 %	-
455265	38	22.5 %	52.7 %	-
501696	12	22.4 %	53.1 %	-
636768	35	22.4 %	53.5 %	-
678375	19	22.4 %	53.8 %	-
443808	27	22.3 %	54.2 %	-
498078	23	22.3 %	54.6 %	-
511344	26	22.2 %	55.0 %	-
637371	20	22.1 %	55.4 %	-
445014	25	22.1 %	55.8 %	-
651240	21	22.0 %	56.2 %	-
431748	22	22.0 %	56.5 %	-
660888	21	22.0 %	56.9 %	-
528831	14	21.9 %	57.3 %	-
446220	19	21.9 %	57.7 %	-
472752	30	21.8 %	58.1 %	-
458883	27	21.8 %	58.5 %	-
463104	16	21.7 %	58.8 %	-
510741	24	21.6 %	59.2 %	-
684405	20	21.6 %	59.6 %	-
626517	31	21.6 %	60.0 %	5 %
529434	34	21.5 %	60.4 %	-
437778	25	21.4 %	60.8 %	-
621693	20	21.3 %	61.2 %	-
665109	12	21.2 %	61.5 %	-
662697	22	21.2 %	61.9 %	-
629532	24	21.2 %	62.3 %	-

Fortsetzung nächste Seite

Übersicht zur Stichprobe (Englisch)

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
459486	29	21.2 %	62.7 %	-
506520	29	21.1 %	63.1 %	-
636165	23	21.1 %	63.5 %	-
426924	16	21.1 %	63.8 %	-
489033	22	21.1 %	64.2 %	-
669330	22	21.0 %	64.6 %	-
434160	19	20.9 %	65.0 %	20 %
613854	41	20.9 %	65.4 %	-
612045	28	20.9 %	65.8 %	-
484209	30	20.7 %	66.2 %	-
528228	21	20.7 %	66.5 %	-
479385	33	20.6 %	66.9 %	-
467928	27	20.5 %	67.3 %	-
614457	23	20.5 %	67.7 %	-
457074	55	20.5 %	68.1 %	-
489636	44	20.4 %	68.5 %	-
642798	18	20.3 %	68.8 %	-
428733	23	20.3 %	69.2 %	-
434763	13	20.2 %	69.6 %	-
497475	42	20.1 %	70.0 %	45 %
651843	19	20.0 %	70.4 %	-
618075	20	20.0 %	70.8 %	-
491445	30	20.0 %	71.2 %	-
664506	34	19.9 %	71.5 %	-
526419	14	19.9 %	71.9 %	-
646416	37	19.9 %	72.3 %	-
630738	24	19.8 %	72.7 %	-
483003	21	19.7 %	73.1 %	-
683199	37	19.7 %	73.5 %	-
532449	12	19.6 %	73.8 %	-
422100	39	19.6 %	74.2 %	-
525213	26	19.6 %	74.6 %	-
692847	30	19.5 %	75.0 %	-
452853	13	19.4 %	75.4 %	-
647019	23	19.0 %	75.8 %	-
454059	29	19.0 %	76.2 %	-
538479	13	18.9 %	76.5 %	-
519183	18	18.9 %	76.9 %	-
513153	17	18.9 %	77.3 %	-
453456	35	18.8 %	77.7 %	-
524007	27	18.8 %	78.1 %	-
469134	19	18.7 %	78.5 %	-
486621	35	18.7 %	78.8 %	-
672948	17	18.7 %	79.2 %	-
431145	49	18.6 %	79.6 %	-

Fortsetzung nächste Seite

Übersicht zur Stichprobe (Englisch)

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
505917	36	18.5 %	80.0 %	0 %
508932	22	18.5 %	80.4 %	-
518580	30	18.5 %	80.8 %	-
444411	22	18.5 %	81.2 %	-
509535	19	18.5 %	81.5 %	-
490842	29	18.4 %	81.9 %	-
466119	19	18.3 %	82.3 %	-
619281	15	18.3 %	82.7 %	-
631944	51	18.1 %	83.1 %	-
469737	31	18.0 %	83.5 %	-
661491	21	18.0 %	83.8 %	-
475164	16	18.0 %	84.2 %	-
427527	28	18.0 %	84.6 %	-
654858	23	18.0 %	85.0 %	-
452250	17	17.9 %	85.4 %	-
423909	22	17.8 %	85.8 %	-
476973	15	17.6 %	86.2 %	-
482400	26	17.4 %	86.5 %	-
493857	19	17.3 %	86.9 %	-
671139	23	17.1 %	87.3 %	-
656667	19	16.8 %	87.7 %	-
460089	18	16.6 %	88.1 %	-
627120	15	16.4 %	88.5 %	-
501093	15	16.4 %	88.8 %	-
446823	50	16.3 %	89.2 %	-
464913	12	16.3 %	89.6 %	-
438984	12	16.2 %	90.0 %	0 %
667521	22	15.9 %	90.4 %	-
530640	22	15.8 %	90.8 %	-
510138	12	15.4 %	91.2 %	-
525816	15	15.0 %	91.5 %	-
582498	15	15.0 %	91.9 %	-
675360	19	15.0 %	92.3 %	-
440190	12	14.9 %	92.7 %	-
527625	15	14.8 %	93.1 %	-
666918	17	14.2 %	93.5 %	-
669933	15	14.1 %	93.8 %	-
499284	21	14.0 %	94.2 %	-
496872	16	13.3 %	94.6 %	0 %
679581	18	12.9 %	95.0 %	-
439587	12	11.5 %	95.4 %	-
485415	13	10.2 %	95.8 %	-
440793	16	8.0 %	96.2 %	-
500490	13	7.5 %	96.5 %	-
479988	12	7.0 %	96.9 %	-

Fortsetzung nächste Seite

Übersicht zur Stichprobe (Englisch)

Segment	Länge	TOP 4 Ähnlichkeit	Rang	Güte
470943	28	6.8 %	97.3 %	-
688626	18	6.5 %	97.7 %	-
639783	26	4.5 %	98.1 %	-
676566	23	4.3 %	98.5 %	-
527022	12	3.0 %	98.8 %	-
668124	16	1.1 %	99.2 %	0 %
492048	24	0.0 %	99.6 %	0 %
623502	16	0.0 %	100.0 %	0 %

Tabelle A.2: Übersicht zur Stichprobe (Englisch)

A.2 Sätze mit hoher Güte der Beleglage

A.2.1 Deutsch

Sätze mit Güte über 60 %. Die Sätze (=Segmente) sind absteigend nach Güte sortiert: 612370-de, 457666-de, 464698-de, 616161-de, 652804-de.

Matches zum Segment 612370-de

Segment 612382, Klasse 'Exact Match' (1.00), $a = 1.0000$

Anfragesatz	Treffer
Da mit dieser übernahme die Voraußetzungen von Artikel 66 Absatz 2 EGKS-Vertrag erfüllt sind , hat die Kommission ihre Zustimmung erteilt	Da mit dieser übernahme die Voraußetzungen von Artikel 66 Absatz 2 EGKS-Vertrag erfüllt sind , hat die Kommission ihre Zustimmung erteilt
Übersetzungsvorschlag	
The proposed transaction is in line with the criteria for the maintenance of competition laid down in Article 66(2) of the ECSC Treaty and may be authorized by the Commission	

Segment 613389, Klasse 'Exact Match' (1.00), $a = 1.0000$

Anfragesatz	Treffer
Da mit dieser übernahme die Voraußetzungen von Artikel 66 Absatz 2 EGKS-Vertrag erfüllt sind , hat die Kommission ihre Zustimmung erteilt	Da mit dieser übernahme die Voraußetzungen von Artikel 66 Absatz 2 EGKS-Vertrag erfüllt sind , hat die Kommission ihre Zustimmung erteilt
Übersetzungsvorschlag	
The proposed transaction is in line with the criteria for the maintenance of competition laid down in Article 66(2) of the ECSC Treaty and was therefore authorized by the Commission	

Segment 625492, Klasse 'nur Tippfehler' (0.95), $a = 0.9928$

Anfragesatz	Treffer
Da mit dieser übernahme die Voraußetzungen von Artikel 66 Absatz 2 EGKS-Vertrag erfüllt sind , hat die Kommission ihre Zustimmung erteilt	Da mit dieser übernahme die Voraußetzungen von Artikel 66 Absatz 2 EGKS-Vertrag erfüllt sind , hat die Kommission ihre Zustimmung erteilt
	.
Übersetzungsvorschlag	
The proposed transaction is in line with the criteria for the maintenance of competition laid down in Article 66 n 2 of the ECSC Treaty and has been authorized by the Commission .	

Segment 632624, Klasse ’nur Tippfehler’ (0.95), $a = 0.7308$

Anfragesatz	Treffer
Da mit	Damit
dieser übernahme die	dieser übernahme die
Vorraußetzungen von Artikel	Voraussetzungen des Artikels
66	66
Absatz	s
2	2
EGKS-Vertrag	EGKS- Vertrag
erfüllt sind , hat die Kommission ihre Zustimmung erteilt	erfüllt sind , hat die Kommission ihre Zustimmung erteilt
	.
Übersetzungsvorschlag	
The proposed transaction is in line with the criteria for the maintenance of competition laid down in Article 66 2 of the ECSC Treaty and has been authorised by the Commission .	

Segment 619902, Klasse ’ähnlicher Inhalt’ (0.70), $a = 0.6087$

Anfragesatz	Treffer
Da mit dieser übernahme	Die Kommission hat ihre Zustimmung erteilt , weil die Prüfung der Anmeldung ergeben hat , daß
die Voraußetzungen von Artikel 66 Absatz 2 EGKS-Vertrag erfüllt sind	die Voraußetzungen von Artikel 66 Absatz 2 EGKS-Vertrag erfüllt sind
, hat die Kommission ihre Zustimmung erteilt	
Übersetzungsvorschlag	
Consideration of the proposed transaction has shown that it meets the tests for authorisation in Article 66 (2) ECSC Treaty and the Commission therefore granted its authorisation	

Segment 613006, Klasse ’ähnlicher Inhalt’ (0.70), $a = 0.5498$

Anfragesatz	Treffer
Da mit dieser übernahme	Die Prüefung der Anmeldung hat ergeben , daß
die Voraußetzungen von Artikel 66	die Voraußetzungen von Artikel 66
Absatz	m
2 EGKS-Vertrag erfüllt sind	2 EGKS-Vertrag erfüllt sind
, hat	;
die Kommission	die Kommission
	hat deshalb
ihre Zustimmung erteilt	ihre Zustimmung erteilt
Übersetzungsvorschlag	
The examination of this transaction has shown that it meets the competition safeguarding tests of Article 66(2) of the ECSC Treaty and the Commission has therefore granted its authorisation	

Segment 625460, Klasse ’ähnlicher Inhalt’ (0.70), $a = 0.5229$

Anfragesatz	Treffer
Da mit dieser übernahme	Die Prüfung der Anmeldung hat ergeben , daß
die Voraußetzungen von Artikel 66	die Voraußetzungen von Artikel 66
Absatz	]
2	2
EGKS-Vertrag	des EGKS-Vertrages
erfüllt sind	erfüllt sind
,	. Die Kommission
hat	hat
die Kommission	deshalb
ihre Zustimmung erteilt	ihre Zustimmung erteilt
Übersetzungsvorschlag	
Consideration of the proposed transaction showed that it meets the tests for authorization in Article 66(2) ECSC Treaty and the Commission therefore granted its authorization	

Segment 625456, Klasse ’ähnlicher Inhalt’ (0.70), $a = 0.5147$

Anfragesatz	Treffer
Da	Da
mit dieser übernahme	das Vorhaben
die Voraußetzungen von Artikel 66	die Voraußetzungen von Artikel 66
Absatz	
2	2
EGKS-Vertrag	des EGKS-Vertrages
erfüllt	erfüllt
sind	
, hat die Kommission ihre	, hat die Kommission ihre
Zustimmung	Genehmigung
erteilt	erteilt
Übersetzungsvorschlag	
The Commission has authorized the transaction since it meets the condi- tions laid down in Article 66(2) of the ECSC Treaty	

Segment 631388, Klasse ’Sub-Segment Match’ (0.60), $a = 0.5950$

Anfragesatz	Treffer
Da mit dieser übernahme	Die Kommission hat diesen Erwerb genehmigt , da
die Voraußetzungen von Artikel 66 Absatz 2 EGKS-Vertrag erfüllt sind	die Voraußetzungen von Artikel 66 Absatz 2 EGKS-Vertrag erfüllt sind
, hat die Kommission ihre Zustimmung erteilt	.
Übersetzungsvorschlag	
Consideration of the proposed transaction has shown that it meets the tests for authorisation in Article 66(2) ECSC Treaty and the Commission therefore granted its authorisation .	

Segment 614934, Klasse 'Sub-Segment Match' (0.60), $a = 0.5281$

Anfragesatz	Treffer
Da	Der Erwerb wurde von der Kommission genehmigt , da
mit dieser übernahme die Voraußetzungen von Artikel 66	mit dieser übernahme die Voraußetzungen von Artikel 66
Absatz	(
2	2
	)
EGKS-Vertrag erfüllt sind	EGKS-Vertrag erfüllt sind
, hat die Kommission ihre Zustimmung erteilt	
Übersetzungsvorschlag	
Consideration of the proposed transaction has shown that it meets the tests for authorization in Article 66(2) of the ECSC Treaty and the Commission therefore granted its authorization	

Matches zum Segment 457666-de

Segment 461809, Klasse 'gleicher Inhalt' (0.90), $a = 0.7755$

Anfragesatz	Treffer
Die	Diese
Aufforderungen	Aufforderungen
ergehen jeweils	erfolgen
in Form einer mit Gründen versehenen Stellungnahme	in Form einer mit Gründen versehenen Stellungnahme
#	
, der zweiten Stufe des Vertragsverletzungsverfahrens	, der zweiten Stufe des Vertragsverletzungsverfahrens
gemäß	nach
Artikel 226 EG-Vertrag .	Artikel 226 EG-Vertrag .
Übersetzungsvorschlag	
These requests take the form of so-called reasoned opinions , the second stage of infringement procedures under Article 226 of the EC Treaty .	

Segment 461790, Klasse 'gleicher Inhalt' (0.90), $a = 0.7062$

Anfragesatz	Treffer
Die	Diese
Aufforderungen	Aufforderungen
ergehen jeweils	erfolgen
in Form	in Form
einer	von
mit Gründen versehenen	mit Gründen versehenen
Stellungnahme #	Stellungnahmen
, der zweiten Stufe des Vertragsverletzungsverfahrens	, der zweiten Stufe des Vertragsverletzungsverfahrens
gemäß	nach
Artikel 226 EG-Vertrag .	Artikel 226 EG-Vertrag .
Übersetzungsvorschlag	
These requests take the form of so-called reasoned opinions , the second stage of infringement procedures under Article 226 of the EC Treaty .	

Segment 470928, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.6374$

Anfragesatz	Treffer
Die Aufforderungen ergehen jeweils	Die Aufforderung erging
in Form einer mit Gründen versehenen Stellungnahme	in Form einer mit Gründen versehenen Stellungnahme
# , der zweiten Stufe des Vertragsverletzungsverfahrens gemäß	im Rahmen des Vertragsverletzungsverfahren nach
Artikel 226 EG-Vertrag .	Artikel 226 EG-Vertrag .
Übersetzungsvorschlag	
The formal request is in the form of a reasoned opinion , under infringement procedures laid down by Article 226 of the EC Treaty .	

Segment 434603, Klasse 'enthält etwas mehr' (0.80), $a = 0.6245$

Anfragesatz	Treffer
Die Aufforderungen ergehen jeweils	Diese formellen Ersuchen werden
in Form einer mit Gründen versehenen Stellungnahme	in Form einer mit Gründen versehenen Stellungnahme
#	abgegeben
, der zweiten Stufe	, der zweiten Stufe
des	eines offiziellen
Vertragsverletzungsverfahrens gemäß Artikel 226 EG-Vertrag .	Vertragsverletzungsverfahrens gemäß Artikel 226 EG-Vertrag .
Übersetzungsvorschlag	
These formal requests will be made in the form of reasoned opinions , the second stage of formal infringement procedures under the EC Treaty (Article 226) . If there is no reply to the reasoned opinion within two months or if the reply is unsatisfactory , the Commission may decide to refer the case to the European Court of Justice .	

Segment 435175, Klasse 'enthält etwas mehr' (0.80), $a = 0.6177$

Anfragesatz	Treffer
Die Aufforderungen ergehen jeweils	Die Aufforderung an Belgien ergeht
in Form einer mit Gründen versehenen Stellungnahme	in Form einer mit Gründen versehenen Stellungnahme
# , der zweiten Stufe des	im Rahmen eines
Vertragsverletzungsverfahrens	Vertragsverletzungsverfahrens
gemäß	nach
Artikel 226 EG-Vertrag .	Artikel 226 EG-Vertrag .
Übersetzungsvorschlag	
The request will be made in the form of a reasoned opinion under the infringement procedure provided for in Article 226 of the Treaty .	

Segment 472951, Klasse 'enthält etwas mehr' (0.80), $a = 0.5697$

Anfragesatz	Treffer
Die Aufforderungen ergehen jeweils	Die formelle Aufforderung Schwedens wird
in Form einer	in Form einer
	sogenannten
mit Gründen	mit Gründen
versehenen	versehene
Stellungnahme #	Stellungnahme #
, der zweiten	(zweite
Stufe des	Stufe des
Vertragsverletzungsverfahrens gemäß	formellen Vertragsverletzungsverfahren nach
Artikel 226 EG-Vertrag	Artikel 226 EG-Vertrag
.	) erfolgen .
Übersetzungsvorschlag	
The formal request to Sweden will take the form of a so-called # reasoned opinion # (second stage of the formal infringement procedure under Article 226 of the EC Treaty) .	

Segment 464651, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.7377$

Anfragesatz	Treffer
Die Aufforderungen ergehen jeweils	Die förmliche Aufforderung ergeht
in Form einer	in Form einer
	so genannten
mit Gründen versehenen Stellungnahme # , der zweiten Stufe des Vertragsverletzungsverfahrens	mit Gründen versehenen Stellungnahme # , der zweiten Stufe des Vertragsverletzungsverfahrens
gemäß	nach
Artikel 226 EG-Vertrag .	Artikel 226 EG-Vertrag .
Übersetzungsvorschlag	
The formal request takes the form of a so-called # reasoned opinion # , the second stage of infringement proceedings under Article 226 of the EC Treaty .	

Segment 455441, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.5958$

Anfragesatz	Treffer
Die Aufforderungen ergehen jeweils in Form einer	Die
mit Gründen	mit Gründen
versehenen	versehene
Stellungnahme	Stellungnahme
# , der zweiten	leitet die zweite
Stufe des Vertragsverletzungsverfahrens gemäß Artikel 226 EG-Vertrag	Stufe des Vertragsverletzungsverfahrens gemäß Artikel 226 EG-Vertrag
.	ein.
Übersetzungsvorschlag	
The sending of a reasoned opinion is the second stage in the infringement procedure provided for in Article 226 of the EC Treaty .	

Segment 459633, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.5898$

Anfragesatz	Treffer
Die Aufforderungen ergehen jeweils in Form	Die Abgabe
einer mit Gründen versehenen Stellungnahme	einer mit Gründen versehenen Stellungnahme
# , der zweiten Stufe	ist das zweite Stadium
des Vertragsverletzungsverfahrens gemäß Artikel 226 EG-Vertrag .	des Vertragsverletzungsverfahrens gemäß Artikel 226 EG-Vertrag .
Übersetzungsvorschlag	
The sending of a Reasoned Opinion is the second stage in the infringement procedure set out in Article 226 of the EC Treaty .	

Segment 434712, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.5728$

Anfragesatz	Treffer
Die Aufforderungen ergehen jeweils in Form	Die Aufforderung durch die Kommission wird im Rahmen
einer	einer
mit Gründen versehenen	begründeten
Stellungnahme	Stellungnahme
#	erfolgen
, der zweiten Stufe des Vertragsverletzungsverfahrens gemäß Artikel 226 EG-Vertrag .	, der zweiten Stufe des Vertragsverletzungsverfahrens gemäß Artikel 226 EG-Vertrag .
Übersetzungsvorschlag	
The Commission will make its request in a reasoned opinion - the second stage in the infringement procedure under Article 226 of the EC Treaty .	

Matches zum Segment 464698-de

Segment 457918, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.6500$

Anfragesatz	Treffer
Die endgültige Entscheidung	Die endgültige Entscheidung
der	über dieses Programmplanungsdokument wird die
Kommission	Kommission
über die beiden Programme wird	
nach	nach
deren	
Prüfung durch den Ausschuss für die Entwicklung und Umstellung der Regionen	Prüfung durch den Ausschuss für die Entwicklung und Umstellung der Regionen
sowie durch	und
den ESF-Ausschuss	den ESF-Ausschuss
ergehen .	erlassen .
Übersetzungsvorschlag	
The final decision on this programming document will be taken by the Commission after they have been considered by the Committee on the Development and Conversion of Regions and the ESF Committee .	

Segment 461083, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.6057$

Anfragesatz	Treffer
Die endgültige Entscheidung der Kommission	Die endgültigen Entscheidungen
über	über
die beiden Programme wird nach	diese Programmplanungsdokumente werden kurz
deren Prüfung durch den Ausschuss für die Entwicklung und Umstellung der	deren Prüfung durch den Ausschuss für die Entwicklung und Umstellung der
Regionen sowie durch	Regionen(1) und
den ESF-Ausschuss ergehen .	den ESF-Ausschuss ergehen .
Übersetzungsvorschlag	
The final decisions on these programming documents will be taken shortly as they have been considered by the Committee on the Development and Conversion of Regions and the ESF Committee .	

Segment 461208, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.6050$

Anfragesatz	Treffer
Die endgültige Entscheidung	Die endgültige Entscheidung
der	über das Programmplanungsdokument trifft die
Kommission	Kommission
über die beiden Programme wird	
nach	nach
deren	
Prüfung durch den Ausschuss für die Entwicklung und Umstellung der	Prüfung durch den Ausschuss für die Entwicklung und Umstellung der
Regionen sowie durch	Regionen(1) und
den ESF-Ausschuss	den ESF-Ausschuss
ergehen .	.
Übersetzungsvorschlag	
The final decision on this programming document will be taken by the Commission after it has been considered by the Committee on the Development and Conversion of Regions and the ESF Committee .	

Segment 461241, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.5939$

Anfragesatz	Treffer
Die endgültige Entscheidung	Die endgültige Entscheidung
	über das Programmplanungsdokument wird von
der Kommission	der Kommission
über die beiden Programme wird	
nach	nach
deren	
Prüfung	Prüfung
	des Programms
durch den Ausschuss für die Entwicklung und Umstellung der	durch den Ausschuss für die Entwicklung und Umstellung der
Regionen sowie durch	Regionen(1) und
den ESF-Ausschuss	den ESF-Ausschuss
ergehen .	getroffen .
Übersetzungsvorschlag	
The final decision on this programming document will be taken shortly as the programme has been considered by the Committee on the Development and Conversion of Regions and the ESF Committee .	

Segment 461103, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.6240$

Anfragesatz	Treffer
Die endgültige Entscheidung	Die endgültige Entscheidung
der Kommission	
über	über
die beiden Programme	das Programmplanungsdokument
wird	wird
	kurz
nach	nach
deren	dessen
Prüfung durch den Ausschuss für die Entwicklung und Umstellung der	Prüfung durch den Ausschuss für die Entwicklung und Umstellung der
Regionen sowie durch	Regionen(1) und
den ESF-Ausschuss ergehen .	den ESF-Ausschuss ergehen .
Übersetzungsvorschlag	
The final decision on this programming document will be taken shortly as the programme has been considered by the Committee on the Development and Conversion of Regions and the ESF Committee .	

Segment 461189, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.5847$

Anfragesatz	Treffer
Die endgültige Entscheidung	Die endgültige Entscheidung
der Kommission	
über	über
die beiden Programme	das Programmplanungsdokument
wird	wird
	kurz
nach	nach
deren Prüfung	dessen Annahme
durch den Ausschuss für die Entwicklung und Umstellung der	durch den Ausschuss für die Entwicklung und Umstellung der
Regionen sowie durch	Regionen(1) und
den ESF-Ausschuss ergehen .	den ESF-Ausschuss ergehen .
Übersetzungsvorschlag	
The final decision on this programming document will be taken shortlyfollowing the approval by the Committee on the Development and Conversion of Regions and the ESF Committee .	

Segment 474379, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.5690$

Anfragesatz	Treffer
Die endgültige Entscheidung	Die endgültige Entscheidung
der	über das Programmplanungsdokument trifft die
Kommission	Kommission
über die beiden Programme wird	
nach	nach
deren	
Prüfung durch den Ausschuss für die Entwicklung und Umstellung der Regionen	Prüfung durch den Ausschuss für die Entwicklung und Umstellung der Regionen
sowie durch den ESF-Ausschuss ergehen .	.
Übersetzungsvorschlag	
The final decision on the programming document will be taken by the Commission after it has been considered by the Committee on the Development and Conversion of the Regions .	

Segment 461062, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.5589$

Anfragesatz	Treffer
Die endgültige Entscheidung	Die endgültige Entscheidung
der Kommission	
über	über
die beiden Programme	das Programmplanungsdokument
wird	wird
	kurz
nach	nach
deren	dessen
Prüfung durch den Ausschuss für die Entwicklung und Umstellung der	Prüfung durch den Ausschuss für die Entwicklung und Umstellung der
Regionen sowie durch den ESF-Ausschuss	Regionen(1)
ergehen .	ergehen .
Übersetzungsvorschlag	
The final decision on this programming document will be taken shortly as the programme after it has been considered by the Committee on the Development and Conversion of Regions .	

Segment 468086, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.5426$

Anfragesatz	Treffer
Die endgültige Entscheidung	Die endgültige Entscheidung
der	über das Programmplanungsdokument trifft die
Kommission	Kommission
über die beiden Programme wird	
nach	nach
deren	dessen
Prüfung durch den Ausschuss für die Entwicklung und Umstellung der	Prüfung durch den Ausschuss für die Entwicklung und Umstellung der
Regionen sowie durch den ESF-Ausschuss ergehen .	Regionen(1) (2) .
Übersetzungsvorschlag	
The final decision on this programming document will be taken by the Commission after it has been considered by the Committee on the Development and Conversion of Regions (1) .	

Segment 472828, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.5426$

Anfragesatz	Treffer
Die endgültige Entscheidung	Die endgültige Entscheidung
der	über das Programmplanungsdokument trifft die
Kommission	Kommission
über die beiden Programme wird	
nach	nach
deren	dessen
Prüfung durch den Ausschuss für die Entwicklung und Umstellung der	Prüfung durch den Ausschuss für die Entwicklung und Umstellung der
Regionen sowie durch den ESF-Ausschuss ergehen .	Regionen(1) (2) .
Übersetzungsvorschlag	
The final decision on this programming document will be taken by the Commission after it has been considered by the Committee on the Development and Conversion of Regions (1) .	

Matches zum Segment 616161-de

Segment 629438, Klasse 'gleicher Inhalt' (0.90), $a = 0.8339$

Anfragesatz	Treffer
Erfolgt keine Antwort auf dieses Schreiben oder ist	Erfolgt keine Antwort auf dieses Schreiben oder ist
diese	die
Antwort nicht überzeugend ,	Antwort nicht überzeugend ,
so geht	unternimmt
die Kommission	die Kommission
zur	den
zweiten	zweiten
Phase über	Schritt
und uebermittelt dem Mitgliedstaat eine mit Gründen versehene Stellungnahme mit der Aufforderung , den festgestellten Verstoß binnen eines Monats abzustellen .	und uebermittelt dem Mitgliedstaat eine mit Gründen versehene Stellungnahme mit der Aufforderung , den festgestellten Verstoß binnen eines Monats abzustellen .
Übersetzungsvorschlag	
The opening of an infringement procedure is formalized by the despatch of a letter of formal notice detailing an alleged failure to comply with Community law and asking the Member State concerned to submit its comments within a month .	

Segment 635882, Klasse 'gleicher Inhalt' (0.90), $a = 0.5287$

Anfragesatz	Treffer
Erfolgt	Ergeht
keine Antwort	keine Antwort
auf dieses Schreiben	,
oder ist	oder ist
diese	die
Antwort nicht	Antwort nicht
überzeugend	ueberzeugend
, so	, so
geht	leitet
die Kommission	die Kommission
zur zweiten	die zweite
Phase	Phase
über und uebermittelt dem Mitgliedstaat	ein , indem sie
eine mit Gründen versehene Stellungnahme	eine mit Gründen versehene Stellungnahme
mit der Aufforderung , den festgestellten Verstoß binnen	abgibt und den Mitgliedstaat auffordert , die festgestellte Vertragsverletzung innerhalb
eines Monats	eines Monats
abzustellen .	aufzuheben .
Übersetzungsvorschlag	
If no reply is received or if the arguments are not convincing , the Commission moves on to the second stage by adopting a reasoned opinion requiring the Member State to terminate the infringement within one month .	

Segment 615173, Klasse ’enthält etwas weniger’ (0.75), $a = 0.4369$

Anfragesatz	Treffer
Erfolgt keine	Wenn eine
Antwort	Antwort
auf dieses Schreiben	ausbleibt
oder	oder
ist diese Antwort nicht überzeugend , so geht	
die Kommission	die Kommission
zur zweiten Phase über und uebermittelt dem Mitgliedstaat	nicht überzeugt , leitet die Kommission die zweite Stufe des Verfahrens ein und gibt
eine mit Gründen versehene Stellungnahme	eine mit Gründen versehene Stellungnahme
mit der Aufforderung , den	ab , um den Mitgliedstaat aufzufordern , dem
festgestellten Verstoß	festgestellten Verstoß
binnen	innerhalb
eines Monats	eines Monats
abzustellen .	ein Ende zu setzen .
Übersetzungsvorschlag	
In the absence of a reply or if the reply is not convincing , the Commission moves on to the second stage and adopts a reasoned opinion enjoining the Member State to put an end to the infringement within a period of one month .	

Segment 615383, Klasse ’ähnlicher Inhalt’ (0.70), $a = 0.4022$

Anfragesatz	Treffer
Erfolgt keine	Wenn eine
Antwort	Antwort
auf dieses Schreiben	ausbleibt
oder	oder
ist diese Antwort nicht überzeugend , so geht	
die Kommission	die Kommission
zur zweiten Phase über und uebermittelt dem Mitgliedstaat	nicht überzeugt , leitet die Kommission die zweite Stufe des Verfahrens ein und gibt
eine mit Gründen versehene Stellungnahme	eine mit Gründen versehene Stellungnahme
mit der Aufforderung , den	ab , um den Mitgliedstaat aufzufordern , dem
festgestellten Verstoß	festgestellten Verstoß
binnen	innerhalb
eines Monats	eines Monats
abzustellen .	ein Ende zu setzen . Andernfalls wird der Gerichtshof der Europäischen Gemeinschaften angerufen .
Übersetzungsvorschlag	
In the absence of a reply or if the reply is not convincing , the Commission moves on to the second stage and adopts a reasoned opinion calling on the Member State to put an end to the infringement within a period of one month .	

Segment 472640, Klasse ’ähnlicher Inhalt’ (0.70), $a = 0.3603$

Anfragesatz	Treffer
Erfolgt keine	Auf der Grundlage der
Antwort	Antwort
auf dieses Schreiben	des Mitgliedstaats (
oder	oder
ist diese	bei einer fehlenden
Antwort	Antwort
nicht überzeugend , so geht	) kann
die Kommission	die Kommission
zur zweiten Phase über und uebermittelt	
dem Mitgliedstaat	dem Mitgliedstaat
	ein zweites Warnschreiben (
eine mit Gründen versehene Stellungnahme	eine mit Gründen versehene Stellungnahme
mit der Aufforderung , den festgestellten	#) übermitteln , in dem sie deutlich die Gründe für den vermuteten
Verstoß	Verstoß
binnen eines Monats abzustellen .	gegen das Gemeinschaftsrecht darlegt und den Mitgliedstaat auffordert , seiner Verpflichtung innerhalb einer bestimmten Frist (im allgemeinen zwei Monate) nachzukommen .
Übersetzungsvorschlag	
In the light of the reply (or absence of a reply) from the Member State concerned , the Commission may decide to address a second written warning (or # Reasoned Opinion #) to the Member State , clearly setting out the reasons why it considers there to have been an infringement of Community law and calling on the Member State to comply within a specified period (normally two months) .	

Segment 466355, Klasse ’ähnlicher Inhalt’ (0.70), $a = 0.3340$

Anfragesatz	Treffer
Erfolgt keine	Nach Eingehen oder Ausbleiben einer
Antwort	Antwort
auf dieses Schreiben oder ist diese Antwort nicht überzeugend , so geht	kann
die Kommission	die Kommission
zur zweiten Phase über und uebermittelt dem Mitgliedstaat eine	beschließen , dem betreffenden Mitgliedstaaten eine #
mit Gründen versehene Stellungnahme	mit Gründen versehene Stellungnahme
mit der Aufforderung , den festgestellten Verstoß binnen	# (zweites Mahnschreiben) zu übermitteln , in der sie klar und eindeutig darlegt , weshalb ihrer Ansicht nach ein Verstoß gegen das Gemeinschaftsrecht vorliegt , und den Mitgliedstaat auffordert , innerhalb
eines	eines
Monats abzustellen .	bestimmten Zeitraums - in der Regel zwei Monaten - dieser Situation abzuhelfen .
Übersetzungsvorschlag	
In the light of the reply or absence of a reply from the Member State concerned , the Commission may decide to address a # Reasoned Opinion # (or second written warning) to the Member State , clearly and definitively setting out the reasons why it considers there to have been an infringement of Community law and calling on the Member State to comply within a specified period (normally two months) , as in this case .	

Segment 461810, Klasse 'Term Match' (0.20), $a = 0.3664$, wird übersprungen
Segment 447998, Klasse 'Term Match' (0.20), $a = 0.3579$, wird übersprungen
Segment 442309, Klasse 'Term Match' (0.20), $a = 0.3445$, wird übersprungen
Segment 472276, Klasse 'Term Match' (0.20), $a = 0.3430$, wird übersprungen

Matches zum Segment 652804-de

Segment 652653, Klasse 'gleicher Inhalt' (0.90), $a = 0.9188$

Anfragesatz	Treffer
Die Kommission hat das Beihilfevorhaben auf die Vereinbarkeit mit dem Beihilfenkodex für die Stahlindustrie hin überprüft und sich vergewissert , daß die Beihilfeintensität sämtlicher Maßnahmen die zulässige Höchstgrenze von 35 % nicht überschreitet und die	Die Kommission hat das Beihilfevorhaben auf die Vereinbarkeit mit dem Beihilfenkodex für die Stahlindustrie hin überprüft und sich vergewissert , daß die Beihilfeintensität sämtlicher Maßnahmen die zulässige Höchstgrenze von 35 % nicht überschreitet und die
in dem	im
Beihilfenkodex für	Beihilfenkodex für
die Stahlindustrie	den Stahlsektor
vorgegebenen Fristen für die Gewährung regionaler Investitionsbeihilfen eingehalten werden .	vorgegebenen Fristen für die Gewährung regionaler Investitionsbeihilfen eingehalten werden .
Übersetzungsvorschlag	
The Commission examined the aid project as to its compatibility with the provisions of the Steel Aids Code , and satisfied itself that the aid intensity of all these measures does not exceed the maximum ceiling of 35 % allowed for , and that the deadlines for granting regional investment aids provided for in the Steel Aids Code will be respected .	

Segment 657933, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.4138$

Anfragesatz	Treffer
Die Kommission hat	Die Kommission hat
das Beihilfevorhaben auf die Vereinbarkeit mit dem Beihilfenkodex für die Stahlindustrie hin überprüft und sich vergewissert , daß	insbesondere geprüft , ob
die Beihilfeintensität sämtlicher	die Beihilfeintensität sämtlicher
Maßnahmen die zulässige Höchstgrenze	Fördermaßnahmen unterhalb der zulässigen Obergrenze
von 35 %	von 35 %
nicht überschreitet und die in dem Beihilfenkodex für die Stahlindustrie vorgegebenen Fristen	liegt , ob mit der Gewährung der Beihilfe ein Abbau der Produktionskapazität in den neuen Bundesländern einhergeht und ob die in dem Stahlbeihilfenkodex
für die Gewährung regionaler Investitionsbeihilfen	für die Gewährung regionaler Investitionsbeihilfen
	vorgesehenen Fristen
eingehalten	eingehalten
werden .	worden sind
Übersetzungsvorschlag	
In particular , the Commission verified that the aid intensity of all the proposed measures remains below the maximum ceiling allowed (35 %) for , that the aid is accompanied by an overall reduction of production capacity in the territory of the former GDR , and satisfied itself that the deadlines for granting regional investment aids , as provided for in the Steel Aids Code , will be respected	

Segment 643332, Klasse ’ähnlicher Inhalt’ (0.70), $a = 0.3953$

Anfragesatz	Treffer
Die Kommission hat	Die Kommission hat
das Beihilfevorhaben auf	
die Vereinbarkeit	die Vereinbarkeit
mit dem Beihilfenkodex für die Stahlindustrie hin überprüft und sich vergewissert	der vorgesehenen Beihilfen mit den Bedingungen des Artikels 5 des Stahlbeihilfenkodex geprüft und insbesondere festgestellt
, daß die	, daß die
Beihilfeintensität sämtlicher Maßnahmen die zulässige Höchstgrenze von 35 % nicht überschreitet und die in dem Beihilfenkodex für die Stahlindustrie vorgegebenen Fristen für die Gewährung regionaler Investitionsbeihilfen eingehalten werden .	Beihilfen im Rahmen von durch die Kommission genehmigten regionalen und allgemeinen Investitionsbeihilferegelungen gewährt werden und die Beihilfeintensität unter der zulässigen Höchstgrenze liegt
Übersetzungsvorschlag	
The Commission has checked that the proposed aid is compatible with Article 5 of the Steel Aid Code and , in particular , has established that the aid is to be granted under the regional and general investment aid schemes approved by the Commission and that the aid intensity does not exceed the ceiling set	

Segment 646149, Klasse ’Term Match’ (0.20), $a = 0.3917$, wird übersprungen
Segment 658887, Klasse ’keine Relevanz’ (0.00), $a = 0.3708$, wird übersprungen
Segment 647600, Klasse ’keine Relevanz’ (0.00), $a = 0.3398$, wird übersprungen
Segment 650343, Klasse ’keine Relevanz’ (0.00), $a = 0.3387$, wird übersprungen
Segment 649865, Klasse ’keine Relevanz’ (0.00), $a = 0.3382$, wird übersprungen
Segment 647518, Klasse ’keine Relevanz’ (0.00), $a = 0.3366$, wird übersprungen
Segment 647722, Klasse ’keine Relevanz’ (0.00), $a = 0.3366$, wird übersprungen

A.2.2 Englisch

Sätze mit Güte über oder gleich 70 %. Die Sätze (=Segmente) sind absteigend nach Güte sortiert: 461295-en, 457677-en, 466722-en, 429939-en, 656064-en, 441396-en.

Matches zum Segment 461295-en

Segment 473824, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.9352$

<table>
<tr><th>Anfragesatz</th><th>Treffer</th></tr>
<tr><td>This # single programming document # amounts to</td><td>This # single programming document # amounts to</td></tr>
<tr><td>518</td><td>98</td></tr>
<tr><td>million in financial support from the European Union .</td><td>million in financial support from the European Union .</td></tr>
<tr><th colspan="2">Übersetzungsvorschlag</th></tr>
<tr><td colspan="2">Die Europäische Kommission hat den Regionalentwicklungsplan des Landes Baden-Württemberg für den Zeitraum 2000-2006 genehmigt . Für dieses # Einheitliche Programmplanungsdokument # werden von der Europäischen Union Fördermittel in Höhe von 98 Mio. bereitgestellt .</td></tr>
</table>

Segment 461211, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.9267$

<table>
<tr><th>Anfragesatz</th><th>Treffer</th></tr>
<tr><td>This # single programming document # amounts to</td><td>This # single programming document # amounts to</td></tr>
<tr><td>518</td><td>808</td></tr>
<tr><td>million in financial support from the European Union .</td><td>million in financial support from the European Union .</td></tr>
<tr><th colspan="2">Übersetzungsvorschlag</th></tr>
<tr><td colspan="2">Für dieses # einheitliche Programmplanungsdokument # werden Fördermittel der Europäischen Union in Höhe von 808 Mio. bereitgestellt .</td></tr>
</table>

Segment 461050, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.9091$

<table>
<tr><th>Anfragesatz</th><th>Treffer</th></tr>
<tr><td>This # single programming document # amounts to</td><td>This # single programming document # amounts to</td></tr>
<tr><td>518</td><td>35.7</td></tr>
<tr><td>million in financial support from the European Union .</td><td>million in financial support from the European Union .</td></tr>
<tr><th colspan="2">Übersetzungsvorschlag</th></tr>
<tr><td colspan="2">Für dieses # einheitliche Programmplanungsdokument # werden Fördermittel der Europäischen Union in Höhe von 35,7 Mio. bereitgestellt .</td></tr>
</table>

Segment 461171, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.9091$

<table>
<tr><th>Anfragesatz</th><th>Treffer</th></tr>
<tr><td>This # single programming document # amounts to</td><td>This # single programming document # amounts to</td></tr>
<tr><td>518</td><td>854</td></tr>
<tr><td>million in financial support from the European Union .</td><td>million in financial support from the European Union .</td></tr>
<tr><th colspan="2">Übersetzungsvorschlag</th></tr>
<tr><td colspan="2">Für dieses # einheitliche Programmplanungsdokument # werden Fördermittel der Europäischen Union in Höhe von 854 Mio. bereitgestellt .</td></tr>
</table>

Segment 461192, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.9091$

Anfragesatz	Treffer
This # single programming document # amounts to	This # single programming document # amounts to
518	189
million in financial support from the European Union .	million in financial support from the European Union .
Übersetzungsvorschlag	
Für dieses # einheitliche Programmplanungsdokument # werden Fördermittel der Europäischen Union in Höhe von 189 Mio. bereitgestellt .	

Segment 473950, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.9091$

Anfragesatz	Treffer
This # single programming document # amounts to	This # single programming document # amounts to
518	171
million in financial support from the European Union .	million in financial support from the European Union .
Übersetzungsvorschlag	
Dieses # einheitliche Programmplanungsdokument # wird von Seiten der Europäischen Union mit Mitteln in Höhe von 171 Mio. finanziell unterstützt .	

Segment 474172, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.9091$

Anfragesatz	Treffer
This # single programming document # amounts to	This # single programming document # amounts to
518	170
million in financial support from the European Union .	million in financial support from the European Union .
Übersetzungsvorschlag	
Dieses # einheitliche Programmplanungsdokument # steht für eine finanzielle Förderung von Seiten der Europäischen Union im Umfang von 170 Millionen EUR .	

Segment 474357, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.9091$

Anfragesatz	Treffer
This # single programming document # amounts to	This # single programming document # amounts to
518	113
million in financial support from the European Union .	million in financial support from the European Union .
Übersetzungsvorschlag	
Dieses # einheitliche Programmplanungsdokument # sieht Fördermittel der Europäischen Union in Höhe von 113 Mio. vor .	

Segment 461125, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.8585$

Anfragesatz	Treffer
This # single programming	This # single programming
	#
document	document
#	
amounts to	amounts to
518	717
million in financial support from the European Union .	million in financial support from the European Union .
Übersetzungsvorschlag	
Für dieses # einheitliche Programmplanungsdokument # werden Fördermittel der Europäischen Union in Höhe von 717 Mio. bereitgestellt .	

Segment 473447, Klasse 'fast gleicher Inhalt' (0.85), $a = 0.7732$

Anfragesatz	Treffer
This # single programming document #	This # single programming document #
amounts to 518	provides to 970
million in financial support from the European Union .	million in financial support from the European Union .
Übersetzungsvorschlag	
Die Europäische Kommission hat den Regionalentwicklungsplan des Landes Nordrhein-Westfalen für den Zeitraum 2000-2006 genehmigt . Für dieses # Einheitliche Programmplanungsdokument # werden von der Europäischen Union Fördermittel in Höhe von 970 Mio. bereitgestellt .	

Matches zum Segment 457677-en

Segment 457682, Klasse 'enthält etwas mehr' (0.80), $a = 0.3273$

Anfragesatz	Treffer
Specialist doctors	Lawyers # freedom to establish
The Commission has sent	The Commission has sent
a	
reasoned	reasoned
opinion to	opinions to Belgium , Spain , France , Ireland , Italy , Luxembourg , the Netherlands and
Portugal requiring that	Portugal requiring that
it notifies	they notify
measures	measures
	taken
to implement Directive	to implement Directive
1999/46/EC .	98/5/EC on the right of lawyers to establish in any EU Member State .
Übersetzungsvorschlag	
Niederlassungsfreiheit von Rechtsanwälten Die Kommission hat Belgien , Spanien , Frankreich , Irland , Italien , Luxemburg , den Niederlanden und Portugal mit Gründen versehene Stellungnahmen übermittelt , in denen sie diese Länder auffordert , die Maßnahmen mitzuteilen , die sie zur Umsetzung der Richtlinie 98/5/EG ergriffen haben . Die Richtlinie betrifft das Recht von Rechtsanwälten , sich in einem beliebigen EU-Mitgliedstaat niederzulassen .	

Segment 457674, Klasse 'enthält etwas mehr' (0.80), $a = 0.3199$

Anfragesatz	Treffer
Specialist doctors	Investor-compensation schemes
The Commission has sent a reasoned opinion to	The Commission has sent a reasoned opinion to
Portugal	the United Kingdom
requiring	requiring
that it notifies	to notify
measures	measures
	taken
to implement	to implement
	, within the territory of Gibraltar ,
Directive	Directive
1999/46/EC .	97/9/EC on investor-compensation schemes (see IP/97/138) .
Übersetzungsvorschlag	
Anlegerentschädigungssysteme Die Kommission hat dem Vereinigten Königreich eine mit Gründen versehene Stellungnahme übermittelt , in der sie dieses auffordert , die Maßnahmen mitzuteilen , die es zur Umsetzung der Richtlinie 97/9/EG über Anlegerentschädigungssysteme (siehe IP/97/138) für das Gebiet von Gibraltar getroffen hat .	

Segment 457671, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.2976$

Anfragesatz	Treffer
Specialist doctors	# Payment systems
The Commission has	The Commission has
sent a	decided to send
reasoned	reasoned
opinion to Portugal requiring that it notifies	opinions to Luxembourg , France and Italy because they have not yet notified the Commission of
measures	measures
	taken
to implement	to implement
	the Settlement Finality
Directive	Directive
1999/46/EC .	(98/26/EC) .
Übersetzungsvorschlag	
# Zahlungssysteme Die Kommission hat beschlossen , Luxemburg , Frankreich und Italien eine mit Gründen versehene Stellungnahme zuzuleiten , da diese Länder der Kommission bislang keine Maßnahmen zur Umsetzung der Richtlinie über die Wirksamkeit von Abrechnungen (Richtlinie 98/26/EG) mitgeteilt haben .	

Segment 442333, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.2874$

Anfragesatz	Treffer
Specialist doctors The Commission has	On 23 December 1999 , the Commission
sent a reasoned opinion to	sent a reasoned opinion to
Portugal requiring that it notifies	Luxembourg in which it requested that the necessary
measures	measures
to implement	be taken in order to transpose
Directive	Directive
1999/46/EC .	96/92/EC into national law .
Übersetzungsvorschlag	
Die Kommission forderte Luxemburg am 23 . Dezember 1999 in einer mit Gründen versehenen Stellungnahme auf , die zur Umsetzung der Richtlinie 96/92/EG in nationales Recht erforderlichen Maßnahmen einzuleiten . Luxemburg antwortete , daß der Entwurf eines Gesetzes zur Umsetzung der Richtlinie 96/92/EG dem Staatsrat vorliege und nach dessen Stellungnahme der Abgeordnetenkammer zur endgültigen Prüfung sowie zur Verabschiedung unterbreitet werden solle .	

Segment 423239, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.2805$

Anfragesatz	Treffer
Specialist doctors	Portugal - incorrect implementation of Services Directive
The Commission has	The Commission has
sent	decided to send
a reasoned opinion to Portugal	a reasoned opinion to Portugal
requiring that it notifies measures	concerning its failure
to implement	to implement
	fully and correctly the
Directive	Directive
1999/46/EC .	on procurement of services (92/50/EEC) .
Übersetzungsvorschlag	
Portugal - Unvorschriftsmäßige Umsetzung der Richtlinie über öffentliche Dienstleistungsaufträge Die Kommission hat beschlossen , wegen der unvollständigen und unkorrekten Umsetzung der Dienstleistungsrichtlinie (92/50/EWG) eine mit Gründen versehene Stellungnahme an Portugal zu richten .	

Segment 430785, Klasse ’Term Match’ (0.20), $a = 0.2965$, wird übersprungen
Segment 426855, Klasse ’Term Match’ (0.20), $a = 0.2893$, wird übersprungen
Segment 423152, Klasse ’Term Match’ (0.20), $a = 0.2767$, wird übersprungen
Segment 430772, Klasse ’keine Relevanz’ (0.00), $a = 0.2841$, wird übersprungen
Segment 427943, Klasse ’keine Relevanz’ (0.00), $a = 0.2748$, wird übersprungen

Matches zum Segment 466722-en

Segment 465386, Klasse ’fast gleicher Inhalt’ (0.85), $a = 0.7516$

Anfragesatz	Treffer
There will be	Provision has been made for
two calls for projects with a view to selecting the development partnerships which will actually be running the schemes .	two calls for projects with a view to selecting the development partnerships which will actually be running the schemes .
Übersetzungsvorschlag	
Für die Auswahl der Entwicklungspartnerschaften zur Durchführung der Maßnahmen sind zwei Aufforderungen zur Einreichung von Vorschlägen vorgesehen .	

Segment 467269, Klasse ’enthält etwas mehr’ (0.80), $a = 0.6506$

Anfragesatz	Treffer
There will be	Programme implementation Provision has been made for
two calls for projects with a view to selecting the development partnerships which will actually be running the schemes .	two calls for projects with a view to selecting the development partnerships which will actually be running the schemes .
Übersetzungsvorschlag	
Durchführung des Programms Vorgesehen sind zwei Aufrufe zur Einreichung von Vorhaben zwecks Auswahl der Entwicklungspartnerschaften , die die Maßnahmen durchführen werden .	

Segment 462898, Klasse ’ähnlicher Inhalt’ (0.70), $a = 0.3388$

Anfragesatz	Treffer
There will be two	Two
calls for	calls for
projects with a view to selecting the development partnerships which will actually be running the schemes .	proposals are scheduled for selecting partnerships for development to implement these activities .
Übersetzungsvorschlag	
Vorgesehen sind zwei Aufrufe zur Einreichung von Vorhaben zwecks Auswahl der Entwicklungspartnerschaften , die die Maßnahmen durchführen werden .	

Segment 465532, Klasse 'Sub-Segment Match' (0.60), $a = 0.3072$

Anfragesatz	Treffer
There will be	The regions are directly responsible for running local schemes (80 % of resources) , while the Ministry of Labour takes on the national coordination of the programme and is directly responsible for running the sectoral schemes (20 % of resources) Provision has been made for
two calls for projects with a view to selecting the development partnerships which will actually be running the schemes .	two calls for projects with a view to selecting the development partnerships which will actually be running the schemes .
Übersetzungsvorschlag	
Zur Auswahl der Entwicklungspartnerschaften , von denen die Maßnahmen durchgeführt werden sollen , sind zwei Aufrufe zur Einreichung von Projektvorschlägen vorgesehen .	

Segment 468591, Klasse 'Term Match' (0.20), $a = 0.3294$, wird übersprungen
Segment 472271, Klasse 'Term Match' (0.20), $a = 0.3294$, wird übersprungen
Segment 463084, Klasse 'Term Match' (0.20), $a = 0.3096$, wird übersprungen
Segment 463119, Klasse 'Term Match' (0.20), $a = 0.3096$, wird übersprungen
Segment 465497, Klasse 'Term Match' (0.20), $a = 0.3096$, wird übersprungen
Segment 647161, Klasse 'keine Relevanz' (0.00), $a = 0.1711$, wird übersprungen

Matches zum Segment 429939-en

Segment 468429, Klasse ’enthält etwas mehr’ (0.80), $a = 0.4378$

Anfragesatz	Treffer
The	However , the Commission considered that the
measures	measures
, however , satisfy the criteria laid down in the European Union	can be approved under the EU rules on State aid and the # Community
guidelines	guidelines
	on State aid
for rescuing and	for rescuing and
restructuring	reestructuring
firms in difficulty	firms in difficulty
and can therefore be approved .	# in particular .
Übersetzungsvorschlag	
Die Kommission kam zu dem Ergebnis , dass diesen Maßnahmen in Anwendung der gemeinschaftlichen Regeln für staatliche Beihilfen und dabei insbesondere der # Gemeinschaftlichen Leitlinien für staatliche Beihilfen zur Rettung und Umstrukturierung von Unternehmen in Schwierigkeiten # zugestimmt werden kann , weil die italienische Regierung nachgewiesen hat , dass sie Bestandteil eines umfassenden Umstrukturierungsplanes sind , mit dem die Rentabilität des Unternehmens innerhalb eines angemessenen Zeitraums unter vorsichtigen Annahmen hinsichtlich der Marktentwicklung wieder hergestellt werden kann .	

Segment 458469, Klasse ’enthält etwas mehr’ (0.80), $a = 0.3930$

Anfragesatz	Treffer
The measures , however , satisfy the criteria	That is why the Commission concluded that the aid granted to Fesa-Enfersa meets the conditions
laid down in the	laid down in the
European Union	Community
guidelines	guidelines
	on State aid
for rescuing and restructuring firms in difficulty and can therefore be	for rescuing and restructuring firms in difficulty and can therefore be
approved .	authorised .
Übersetzungsvorschlag	
Deshalb ist die Kommission zu dem Schluss gekommen , dass die Fesa-Enfersa gewährte Finanzhilfe in Einklang steht mit den in den Leitlinien der Gemeinschaft für staatliche Beihilfen zur Rettung und Umstrukturierung von Unternehmen in Schwierigkeiten festgelegten Bedingungen und daher genehmigt werden kann .	

Segment 459838, Klasse ’ähnlicher Inhalt’ (0.70), $a = 0.3615$

Anfragesatz	Treffer
The measures , however , satisfy the criteria laid down in the European Union	The Commission approved the aid since it found the aid to be compatible with the 1994 Community
guidelines	guidelines
	on State aid
for rescuing and restructuring firms in difficulty	for rescuing and restructuring firms in difficulty
and can therefore be approved .	.
Übersetzungsvorschlag	
Die Kommission genehmigte die Beihilfen gemäß den gemeinschaftlichen Beihilfen für staatliche Beihilfen zur Rettung und Umstrukturierung von Unternehmen in Schwierigkeiten aus dem Jahr 1994 .	

Segment 471504, Klasse 'Sub-Segment Match' (0.60), $a = 0.3772$

Anfragesatz	Treffer
The measures , however , satisfy the criteria laid down in the European Union	The Commission considered the aid to be compatible with the Treaty(2) and with the
guidelines for rescuing and restructuring firms in difficulty	guidelines for rescuing and restructuring firms in difficulty
and can therefore be approved .	.
Übersetzungsvorschlag	
Nach Auffassung der Kommission steht die Beihilfe mit den Bestimmungen des EG-Vertrags(2) und den Leitlinien der Gemeinschaft für staatliche Beihilfen zur Rettung und Umstrukturierung von Unternehmen in Schwierigkeiten im Einklang . Die Beihilfe beeinträchtige die Handelsbedingungen in keiner dem gemeinsamen Interesse zuwiderlaufenden Weise und erfülle die folgenden Kriterien(3) : Die Firma # Sernam # entspricht der Definition eines Unternehmens in Schwierigkeiten . Die Übernahme durch das Unternehmen # Geodis # und der durchzuführende Umstrukturierungsplan dürften eine Wiederherstellung der Rentabilität innerhalb von vier Jahren ermöglichen . Auf expandierenden Märkten mit hohen Wachstumsraten bedeutet die Beihilfe keine Beeinträchtigung des Wettbewerbs , insbesondere aufgrund des beabsichtigten Abbaus von Arbeitsplätzen und Produktionskapazitäten .	

Segment 425367, Klasse 'Term Match' (0.20), $a = 0.4338$, wird übersprungen
Segment 454689, Klasse 'Term Match' (0.20), $a = 0.3935$, wird übersprungen
Segment 451139, Klasse 'Term Match' (0.20), $a = 0.3884$, wird übersprungen
Segment 465134, Klasse 'Term Match' (0.20), $a = 0.3693$, wird übersprungen
Segment 441489, Klasse 'Term Match' (0.20), $a = 0.3682$, wird übersprungen
Segment 424594, Klasse 'Term Match' (0.20), $a = 0.3613$, wird übersprungen

Matches zum Segment 656064-en

Segment 655225, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.6827$

Anfragesatz	Treffer
The project will be implemented by	The project will be implemented by
	Spanish Médecins Sans Frontières , an
ECHO	ECHO
#s	
partner	partner
, Médecins sans Frontières- Netherlands .	.
Übersetzungsvorschlag	
Diese Hilfe wird über den ECHO-Partner Médecins Sans Frontières - Spanien abgewickelt .	

Segment 656051, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.6697$

Anfragesatz	Treffer
The project will be implemented by ECHO #s	The project will be implemented by ECHO #s
	operational
partner , Médecins	partner , Médecins
sans Frontières- Netherlands .	Sans Frontières Belgium .
Übersetzungsvorschlag	
Dieses Hilfsmaßnahmen werden von der NRO Médecins Sans Frontières-Belgique durchgeführt , die mit dem Amt für humanitäre Hilfen der Europäischen Gemeinschaft zusammenarbeitet .	

Segment 657916, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.6342$

Anfragesatz	Treffer
The project	Some ECU 500 000
will be implemented by ECHO #s partner	will be implemented by ECHO #s partner
	in the operation
, Médecins sans	, Médecins sans
Frontières- Netherlands .	Frontières-Netherlands .
Übersetzungsvorschlag	
Médecins Sans Frontières , Niederlande , ist der Partner des Europäischen Amtes für humanitäre Hilfe (ECHO) , der das mit 500.000 ECU bezifferte Projekt abwickelt .	

Segment 656264, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.5455$

Anfragesatz	Treffer
The project will be implemented by ECHO #s	The project will be implemented by ECHO #s
	NGO
partner , Médecins	partner , Médecins
sans Frontières- Netherlands .	du Monde-France
Übersetzungsvorschlag	
Durchgeführt wird das Projekt von der französischen NRO # Médecins du Monde # , einem Partner von ECHO	

Segment 655217, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.4635$

Anfragesatz	Treffer
The	This
project will be implemented by	project will be implemented by
	a number of
ECHO #s	ECHO #s
partner	NGO partners , including the Spanish Red Cross
, Médecins	, Médecins
sans Frontières- Netherlands .	Sans Frontières and Médecins du Monde .
Übersetzungsvorschlag	
Diese Aktion wird von mehreren NRO durchgeführt mit denen ECHO zusammenarbeitet , darunter die spanischen Abteilungen des Roten Kreuzes , von Médecins Sans Frontières und von Médecins du Monde .	

Segment 614322, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.4286$

Anfragesatz	Treffer
The project	It
will be implemented by	will be implemented by
ECHO #s partner ,	the Belgian branch of the NGO
Médecins sans	Médecins sans
Frontières- Netherlands .	Frontières
Übersetzungsvorschlag	
Die Abwicklung besorgt der belgische Zweig der nichtstaatlichen Organisation Médecins sans frontières	

Segment 657925, Klasse ’Sub-Segment Match’ (0.60), $a = 0.5953$

Anfragesatz	Treffer
The	The six-month
project will be implemented by ECHO #s partner	project will be implemented by ECHO #s partner
	in the operation
, Médecins sans	, Médecins sans
Frontières- Netherlands .	Frontières-France .
Übersetzungsvorschlag	
Das Projekt , das eine Laufzeit von sechs Monaten hat , soll von Médecins Sans Frontières , Frankreich , als operationellem Partner des Europäischen Amtes für humanitäre Hilfe abgewickelt werden .	

Segment 622264, Klasse ’Term Match’ (0.20), $a = 0.4415$, wird übersprungen
Segment 656072, Klasse ’Term Match’ (0.20), $a = 0.4228$, wird übersprungen
Segment 616675, Klasse ’Term Match’ (0.20), $a = 0.3778$, wird übersprungen

Matches zum Segment 441396-en

Segment 456281, Klasse ’ähnlicher Inhalt’ (0.70), $a = 0.4530$

Anfragesatz	Treffer
However , Commission	The Commission #s
investigations	investigations
	have
revealed that the market position of the	revealed that the market position of the
two firms posed no likelihood	parties precludes the emergence or strengthening
of a dominant position	of a dominant position
being created or strengthened .	on the market .
Übersetzungsvorschlag	
Die Untersuchungen der Kommission haben ergeben , dass die Marktstellung der Parteien die Entstehung oder Verstärkung von Marktbeherrschung ausschließt .	

Segment 455823, Klasse ’ähnlicher Inhalt’ (0.70), $a = 0.4338$

Anfragesatz	Treffer
However , Commission	The Commission #s
investigations	investigations
revealed that the	have established that the parties # positions on the
market	market
position of the two firms posed no likelihood	preclude the creation or strengthening
of a dominant position	of a dominant position
being created or strengthened .	.
Übersetzungsvorschlag	
Die Untersuchungen der Kommission haben ergeben , dass die Marktstellung der Parteien die Entstehung oder Verstärkung von Marktbeherrschung ausschließt .	

Segment 431439, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.4337$

<table>
<tr><th>Anfragesatz</th><th>Treffer</th></tr>
<tr><td>However , Commission investigations revealed that the market position of the two firms posed no likelihood of</td><td>The Commission #s investigation showed that the operation will not create or strengthen</td></tr>
<tr><td>a dominant position</td><td>a dominant position</td></tr>
<tr><td>being created or strengthened .</td><td>on the market .</td></tr>
<tr><th colspan="2">Übersetzungsvorschlag</th></tr>
<tr><td colspan="2">Die Untersuchung der Kommission hat ausgeschlossen , dass das Vorhaben zur Begründung oder Verstärkung einer marktbeherrschenden Stellung führen wird .</td></tr>
</table>

Segment 432391, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.3947$

<table>
<tr><th>Anfragesatz</th><th>Treffer</th></tr>
<tr><td>However , Commission investigations revealed that the market position of the two firms posed no likelihood of</td><td>The Commission #s investigation led it to conclude that the operation is not likely to create or strengthen</td></tr>
<tr><td>a dominant position</td><td>a dominant position</td></tr>
<tr><td>being created or strengthened .</td><td>on the relevant markets .</td></tr>
<tr><th colspan="2">Übersetzungsvorschlag</th></tr>
<tr><td colspan="2">Die Untersuchung der Kommission hat ergeben , dass das Vorhaben nicht geeignet ist , auf den relevanten Märkten eine beherrschende Stellung zu begründen oder zu verstärken .</td></tr>
</table>

Segment 435491, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.3687$

<table>
<tr><th>Anfragesatz</th><th>Treffer</th></tr>
<tr><td>However , Commission investigations revealed that the market position of the two firms posed no likelihood</td><td>The Commission #s investigation showed that the proposed concentration will not lead to the creation</td></tr>
<tr><td>of a dominant position</td><td>of a dominant position</td></tr>
<tr><td>being created or strengthened .</td><td>on the market .</td></tr>
<tr><th colspan="2">Übersetzungsvorschlag</th></tr>
<tr><td colspan="2">Der Prüfung der Kommission zufolge wird dieser Zusammenschluß keine beherrschende Marktposition bewirken .</td></tr>
</table>

Segment 436847, Klasse 'ähnlicher Inhalt' (0.70), $a = 0.3512$

<table>
<tr><th>Anfragesatz</th><th>Treffer</th></tr>
<tr><td>However , Commission investigations revealed that the market position of the two firms posed no likelihood of</td><td>Following investigations , the Commission concluded that the transaction will not create or reinforce</td></tr>
<tr><td>a dominant position</td><td>a dominant position</td></tr>
<tr><td>being created or strengthened .</td><td>on the relevant markets .</td></tr>
<tr><th colspan="2">Übersetzungsvorschlag</th></tr>
<tr><td colspan="2">Die von der Kommission vorgenommene Prüfung lässt den Schluss zu , dass die Maßnahme nicht zu einer Begründung oder Verstärkung einer beherrschenden Stellung auf den fraglichen Märkten führt .</td></tr>
</table>

Segment 441392, Klasse 'Term Match' (0.20), $a = 0.3618$, wird übersprungen
Segment 440091, Klasse 'Term Match' (0.20), $a = 0.3591$, wird übersprungen
Segment 431923, Klasse 'Term Match' (0.20), $a = 0.3512$, wird übersprungen
Segment 652088, Klasse 'Term Match' (0.20), $a = 0.3483$, wird übersprungen

Literaturverzeichnis

ALESIANI, EMILIO (1997): “Considerations in Open Translation Memory”. *The LISA Newsletter* XI (3.6). Online verfügbar.

BALDWIN, TIMOTHY UND TANAKA, HOZUMI (2000): “The Effects of Word Order and Segmentation on Translation Retrieval Performance”. In: *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. Saabruecken, S. 35–41. PDF online verfügbar.

BOWKER, LYNNE (1998): “Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study”. *META* XLIII 4.

BOWKER, LYNNE (2002): *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press. ISBN 0-7766-3016-4.

BRANTS, THORSTEN (2000): “TnT - A Statistical Part-of-Speech Tagger”. In: *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*. Seattle, S. 224–231.

CARL, MICHAEL UND HANSEN, SILVIA (1999): “Linking Translation Memories with Example-Based Machine Translation”. Technischer Bericht 36. IAI Working Paper.

CARL, MICHAEL UND WAY, ANDY (2003): “Introduction”. In: *Recent Advances in Example-Based Machine Translation*, herausgegeben von Carl, Michael und Way, Andy, Kluwer Academic Publishers.

CARLETTA, JEAN (1996): “Assessing agreement on classification tasks: the kappa statistics”. *Computational Linguistics* 22 (2): S. 249–254.

CRANIAS, LAMBROS; PAPAGEORGIOU, HARRIS UND PIPERIDIS, STELIOS (1994): “A Matching Technique in Example-Based Machine Translation”. In: *Coling*. S. 100–104.

DENNETT, GERALD (1995): “Translation Memory: Concept, products, impact and prospects”. project report, South Bank University.

ERJAVEC, TOMAŽ (1999): “The ELAN Slovene-English Aligned Corpus”. In: *Proceedings of the Machine Translation Summit VII*. S. 349–357.

ERPENBECK, ARNO; HELLMANN, DANIELA; PETERS, TONY; SCHMEIER, FRAUKE; STEFFENS, TIMO; SURREY, ANNIKA UND WAGNER, JOACHIM (2000): “Translation Memory”. Seminararbeit. `http://www-lehre.informatik.uni-osnabrueck.de/~jwagner/tm/`.

ERPENBECK, ARNO; KOCH, BRITTA; KUMMER, NORMAN; REUTER, PHILIP; TSCHORN, PATRICK UND WAGNER, JOACHIM (2002): “KOKS – Korpusbasierte Kollokationssuche”. Technischer Bericht, Institut für Kognitionswissenschaft, Universität Osnabrück. Abschlussbericht.

FEDER, MARCIN (2001): *Computer Assisted Translation. A Proposal for Tool Evaluation Methodology*. Dissertation, Adam Mickiewicz University, Poznań, Polen. Bibliographie online verfügbar.

GHORBEL, HATEM; CORAY, GIOVANNI; LINDEN, ANDRÉ; COLLET, OLIVIER UND AZZAM, WAGIH (2002): “L’alignement multicritères des documents médiévaux”. *Lexicometrica* Numéro spécial: Corpus alignés.

KUMMER, NORMAN UND WAGNER, JOACHIM (2002): “Phrase processing for detecting collocations with KoKS”. Workshop on Computational Approaches to Collocations. `http://www.ai.univie.ac.at/colloc02/kummer_wagner_final.pdf`.

LEECH, G. UND SMITH, N. (1999): “The Use of Tagging”. In: *Syntactic Wordclass Tagging*, herausgegeben von van Halteren, Hans, Kluwer Academic Publishers, S. 23–36.

MANNING, CHRISTOPHER D. UND SCHÜTZE, HINRICH (1999): *Foundations of statistical natural language processing*. Cambridge, MA, London: MIT Press.

MCTAIT, KEVIN (2001): “Memory-Based Translation Using Translation Patterns”. In: *Proceedings of the 4th Annual CLUK Colloquium*. Sheffield, S. 43–52.

MELBY, ALAN (1998): “Data exchange standards from the OSCAR and MARTIF projects”. In: *First International Conference on Language Resources and Evaluation, LREC 98*. ELRA, Granada, S. 3–8.

MERKEL, MAGNUS (2001): “Comparing source and target texts in a translation corpus.” 13th Nordic Conference on Computational Linguistics, NoDaLiDa’01. `http://www.ida.liu.se/~magme/publications/merkel-comparing.pdf`.

PIPERIDIS, STELIOS; PAPAGEORGIOU, HARRIS UND BOUTSIS, SOTIRIS (2000): “From sentences to words and clauses”. In: *Parallel Text Processing. Alignment and Use of Translation Corpora*, herausgegeben von Véronis, Jean, Kluwer, S. 117–138.

PLANAS, EMMANUEL UND FURUSE, OSAMU (2000): “Multi-level Similar Segment Matching Algorithm for Translation Memories and Example-Based Machine Translation”. In: *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. Saarbrücken, S. 35–41.

REINKE, UWE (1999): “Evaluierung der linguistischen Leistungsfähigkeit von Translation Memory-Systemen”. *LDV Forum* (16): S. 100–117.

SARDINHA, ANTONIO PAULO BERBER (1997): *Automatic Identification of Segments in Written Text*. Dissertation, University of Liverpool.

SCHMID, HELMUT (1994): “Probabilistic Part-of-Speech Tagging using Decision Trees”. überarbeitete Online-Fassung verwendet. `http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf`.

SCHMID, HELMUT (1995): "Improvements in Part-of-Speech Tagging with an Application to German". überarbeitete Online-Fassung verwendet. http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.pdf.

SEEWALD-HEEG, UTA UND NÜBEL, RITA (1999): "Ausblick". *LDV Forum* (16): S. 118–121.

SIMARD, MICHEL UND LANGLAIS, PHILIPPE (2001): "Sub-sentential exploitation of translation memories". In: *Proceedings of MT Summit VIII*. Santiago de Compostela, Spanien.

SOMERS, HAROLD (1999): "Review Article: Example-based Machine Translation". *Machine Translation* 14 (2): S. 113–158.

SOMERS, HAROLD; MCLEAN, IAN UND JONES, DANIEL (1994): "Experiments in Multilingual Example-Based Generation". In: *Proceedings of the 3rd Conference on the Cognitive Science of Natural Language Processing*. Dublin.

TOUTANOVA, KRISTINA; KLEIN, DAN; MANNING, CHRISTOPHER D. UND SINGER, YORAM (2003): "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network". HLT-NAACL 2003. http://nlp.stanford.edu/˜manning/papers/tagging.pdf.

TSCHORN, PATRICK (2002): *Automatically aligning English-German parallel texts at sentence level using linguistic knowledge*. Magisterarbeit, Universität Osnabrück.

UNBEKANNT (2001): "Ohne Titel". Laut Language Automation, Inc. handelt es sich um ein von Trados bereitgestelltes Dokument, das von SDL und Brian Chandler (MultiLing Corp.) aktualisiert wurde. http://www.lai.com/tmcompet.htm.

VAN HALTEREN, HANS UND VOUTILAINEN, ATRO (1999): "Automatic Taggers: An Introduction". In: *Syntactic Wordclass Tagging*, herausgegeben von van Halteren, Hans, Kluwer Academic Publishers, S. 109–115.

VÉRONIS, JEAN (Herausgeber) (2000): *Parallel Text Processing. Alignment and Use of Translation Corpora*. Dordrecht, Niederlande: Kluwer. ISBN 0-7923-6546-1.

WEBB, LYNN E. (1998): *Advantages and Disadvantages of Translation Memory: A Cost/Benefit Analysis*. Magisterarbeit, Monterey Institute of International Studies (MIIS), Monterey, Kalifornien. Die online verfügbaren Fassungen haben unterschiedliche Seitenbreiten und -nummerierungen.

WIBLE, DAVID; YI CHIEN, FENG; KUO, CHIN-HWA UND WANG, CC (2002): "Towards Automating a Personalized Concordancer for Data-Driven Learning: A Lexical Difficulty Filter for Language Learners". In: *Teaching and Learning by Doing Corpus Analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora, Graz*, herausgegeben von Kettemann, Bernhard und Marko, Georg. Amsterdam – New York: Rodopi.

Der nachfolgende Index enthält zwar nicht alle Namen und Begriffe, die im Text auftreten. Er wurde aber beibehalten, da er umfangreich genung ist, um eine Hilfe sein zu können. Der Autor bittet um Nachsicht, wenn wichtige Begriffe nicht gefunden werden können.

Index

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Dublin, den 11. September 2003